Applications of Genetic Algorithms on the Structure−Activity Relationship Analysis of Some Cinnamamides

T. J. Hou, J. M. Wang, N. Liao, and X. J. Xu*

Department of Chemistry, Peking University, Beijing 100871, People's Republic of China

Received February 14, 1999

Quantitative structure−activity relationships (QSARs) for 35 cinnamamides were studied. By using a genetic algorithm (GA), a group of multiple regression models with high fitness scores was generated. From the statistical analyses of the descriptors used in the evolution procedure, the principal features affecting the anticonvulsant activity were found. The significant descriptors include the partition coefficient, the molar refraction, the Hammet σ constant of the substituents on the benzene ring, and the formation energy of the molecules. It could be found that the steric complementarity and the hydrophobic interaction between the inhibitors and the receptor were very important to the biological activity, while the contribution of the electronic effect was not so obvious. Moreover, by construction of the spline models for these four principal descriptors, the effective range for each descriptor was identified.

INTRODUCTION

3,4-(Methylenedioxy)cinnamoyl piperidide with instinctive anticonvulsant activity, simplified from poperine II, has been identified as a potential antieplilepsy drug.[1] Clinical use showed that this compound actually had relatively good therapeutic effectiveness for different epileptic patients and relatively few untold effects. The basic structure of 3,4-(methylenedioxy)cinnamoyl piperidide (see Chart 1) is that of a vinyl group (B region), with a hydrophobic benzene ring on one end (A region) and an amido group (C region) on the other. Our previous study showed that the benzene ring of part A was necessary for the activity. On the benzene ring, the substitution of 4-chloro groups, 2-chloro groups, and so on for hydrogen atoms would increase the anticonvulsant activity. The presence of the −CH═CH− group in part B was also important. When the double bond was saturated or shortened to one carbon atom, the anticonvulsant activity would be remarkably reduced. In part C, the amides composed of the amines of relatively small groups, e.g., isopropylamine, *sec*-butylamine, and cycloamylamine, showed stronger anticonvulsant activity than the others.[2]

As is well-known, the cinnamamide analogues (Chart 2) had a wide spectrum of physiological functions,[3−5] including nervous suppression, hypnosis, sedation, anticonvulsion, muscular relaxation, local anesthesia, mycostate etc. Until now, however, very few studies on the relationship between the chemical structures and the biological functions of these kinds of compounds had been reported. Extensive studies about the anticonvulsive activity of this group of compounds have been performed in our laboratory. Early work by us had established a structure−activity profile only for a small set of cinnamamide analogues.[6] Now, a more profound correlation study was accomplished after synthesis of many new compounds. The study was expected to provide insight into the anticonvulsant mechanisms of the cinnamamides and give some useful information that could help researchers design new candidates as potential drugs.

Chart 1. Structure of 3,4-(Methylenedioxy)cinnamoyl Piperidide

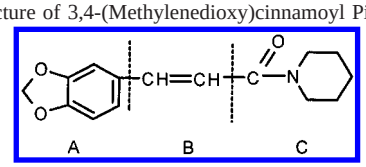
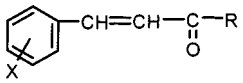


Chart 2. General Structure of the Cinnamamide Derivatives



However, from the experimental viewpoint, the mechanism of the actions of the cinnamamides was not yet properly understood. In the drug−receptor recognition process, the electronic, steric, and hydrophobic characters of the molecules may be important factors affecting the biological activity. Electronic and steric characters of the molecules play an important role in the drug−receptor interaction, because they directly affect the energetic complementarity and the steric complementarity of interaction molecules at the active sites. At the same time, hydrophobicity should be considered because the hydrophobic interactions of drugs in the biological system are often very significant. So a correlation study based on a wide variety of molecular descriptors was expected to provide insight into the anticonvulsant mechanisms of cinnamamides.

To take full consideration of the effects of electronic, steric, and hydrophobic features in drug-affecting processes, a total of 19 descriptors representing these factors was used. A quantitative structure−activity relationship (QSAR) analysis method based on a GA had been developed in our laboratory[7] and was applied in this QSAR study. The goal of this study was to develop QSARs for cinnamamides and determine whether the obtained descriptors can help us understand the biological activity of the drugs in this

* To whom correspondence should be addressed.

10.1021/ci990010n CCC: $18.00
Published on Web 08/19/1999

Table 1. Structures of Cinnamamide Derivatives[a] and Experimental and Calculated Biological Activities by Eqs 4 and 18

no.	R	X	log(1/*C*) obsd	log(1/*C*) calcd[c]	residue[c]	log(1/*C*) calcd[d]	residue[d]
1		3-Cl	0.788	0.510	0.278	0.595	0.193
2		3-F	0.578	0.500	0.078	0.561	0.017
3		4-F	0.458	0.501	−0.043	0.458	0.000
4		4-Br	0.314	0.442	−0.128	0.500	−0.186
5		2,4-Cl	0.664	0.651	0.013	0.623	0.041
6[b]		3,4-Cl	0.550	0.647	−0.097	0.621	−0.071
7		4-Cl	0.606	0.514	0.092	0.596	0.010
8		4-NO_2	0.268	0.324	−0.056	0.314	−0.046
9		3-NO_2	0.324	0.323	0.001	0.310	0.014
10[b]		3-CF_3	0.921	0.815	0.106	0.899	0.022
11		2-CF_3	0.723	0.797	−0.074	0.899	−0.176
12		4-CF_3	0.921	0.819	0.102	0.899	0.022
13		3-OH, 4-OCH_3	−0.272	−0.237	−0.035	−0.272	0.000
14		4-OCH_3	0.218	0.174	0.044	0.270	−0.052
15		3-I	0.320	0.472	−0.152	0.390	−0.070
16		4-OC_2H_5	0.500	0.242	0.258	0.360	0.140
17[b]		4-OC_3H_7-*n*	0.290	0.332	−0.042	0.348	−0.058
18		4-OC_4H_9-*n*	0.180	0.400	−0.220	0.268	−0.088
19		3-Cl	0.410	0.586	−0.176	0.651	−0.241
20		3-F	0.495	0.573	−0.078	0.366	0.129
21		4-F	0.495	0.574	−0.079	0.561	−0.066
22		4-Br	0.540	0.517	0.023	0.557	−0.017
23[b]	−NHC_4H_9-*i*	2,4-Cl_2	0.735	0.732	0.003	0.684	0.051
24		3,4-Cl_2	0.977	0.718	0.259	0.779	0.198
25		4-Cl	0.714	0.583	0.134	0.653	0.061
26		4-CF_3	0.772	0.892	−0.120	0.899	−0.127
27		3-CF_3	0.989	0.890	0.099	0.899	0.090
28[b]		3-Cl	0.620	0.570	0.050	0.600	0.020
29		4-F	0.288	0.562	−0.274	0.366	−0.078
30		4-Br	0.580	0.507	0.073	0.506	0.074
31	−NHC_3H_7-*i*	2,4-Cl_2	0.600	0.709	−0.109	0.634	−0.034
32		4-Cl	0.801	0.572	0.229	0.603	0.198
33		3,4-Cl_2	0.498	0.704	−0.206	0.629	−0.131
34		4-CF_3	0.899	0.874	0.025	0.899	0.000
35		3-CF_3	0.924	0.875	−0.047	0.899	0.025

[a] See Chart 2. [b] These compounds were used as the test set and are not included in the derivation of equations. [c] The values of log(1/*C*) were calculated using eq 4. [d] The values of log(1/*C*) were calculated using eq 18.

category. On the basis of best QSAR model obtained, we expected to find more potential compounds with the aid of the computational combinatorial chemistry method.

MATERIALS

Experimental Data. Thirty-five cinnamamide analogues were synthesized (see Table 1).[8] The chemical structures of these compounds were all modified from 3,4-(methylenedioxy)cinnamoyl piperidide. These compounds were tested on mice for anticonvulsant activity through maximum electroshock seizure tests (MES), and the value of ED_{50} could be calculated by using the Weil method.[8] The potency was defined as log(1/*C*) (*C* represents ED_{50}) in the QSAR analysis and was used as a dependent variable in the QSAR study (see Table 1).

Calculation of the Molecular Physicochemical Properties. The molecular geometries of all compounds in Table 1 were modeled using the InsightII molecular simulation software package.[9] The initial structures were first minimized using molecular mechanics with consistent-valence force field (CVFF).[10] For some relatively flexible structures, conformational analyses were performed to find their lowest-energy conformers. Then these structures were fully optimized, and some quantum-chemical features were calculated based on the semiempirical AM1 method, available in MOPAC 7.0 on a PC.[11] Partition coefficients were measured by using the method proposed by Hansch.[8] The aqueous desolvation free energy was calculated from the hydration shell model developed by Hopfinger,[12] and the molar refraction came directly from ref 2.

METHODS

QSARs Based on GAs. Recently, some published papers suggested that genetic algorithms (GAs) might be useful in data analysis, especially in the task of reducing the number of features for regression models.[13−15] Rogers and Hopfinger first applied this method in QSAR analysis[15] and proved GA a very effective tool and had many merits that other methods did not have. Compared with other traditional statistical methods, QSARs based on GAs used many models and tested only the final, fully constructed models. GA-based QSARs not only could find a group of reliable QSAR models from a large number of samples but also could construct higher-order polynomials, splines, and Gaussian models. Moreover, from the analyses of the variables used in the evolution procedure, we might obtain the crucial physicochemical properties related to the activity.

The QSAR based on the GA analysis program used in this study was under development in our laboratory and had been embedded into the Peking University Drug Design System as a separate module.

The brief basic steps of the QSAR based on a GA are involved and are as follows.

a. 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 a series of descriptors are randomly chosen as a string. Every descriptor is expressed using two digits; one digit represents its serial number, and the other represents its function type. The initial population is generated by randomly selecting some number of descriptors from the training set. Then these individuals are scored according to their fitness score. An elite population is used to retain the best and different individuals.

b. Crossover Operation. Once all models in the population have been rated using the fitness score, the crossover operation is performed repeatedly. In the operation, two good models are probabilistically selected as "parents" with the likelihood of being chosen proportional to a model fitness score; a pair of children are produced by dividing both parents at a randomly chosen point and then joining the pieces together.

c. Mutation Operation. After crossover operation, mutation operation may randomly alter all individuals in the new population, and the new model fitness is determined.

d. Comparison Operation. After the crossover and mutation operation, the newly created population and the elite population are compared. If there are some individuals in the newly created population that are better than some individuals in the elite population, these better individuals are copied to the elite population. When the total fitness of the elite population cannot be improved, "convergence" is achieved.

e. Partial Reinitialization. A partial reinitialization procedure is easily introduced into the genetic algorithm by replacing the lowest 50−80% chromosomes in the population with randomly generated ones after several steps of crossover and mutation operations. Thus, the likelihood of the GA converging on a local−optimal minimum is reduced. Generally, three to six reinitializations are enough to find all the different QSAR models.

Upon completion, from the elite population, the models with the highest fitness scores can be obtained. For a population of 200 models, if the data set contains 20 features, 500−1000 cycles are usually sufficient to achieve convergence, while 1000−1500 operations are enough when the data set has 30 features. For a typical data set, this process takes 10 min to 1 h on a PC (Pentium 150).

Reliability of the Models. The models in the elite population were sorted by their fitness scores. In this study, the fitness function was defined as the multiple linear regression coefficient (r). The reliabilities of the models were mainly tested with their leave-one-out cross-validated correlation coefficient (Q^2) scores and their actual predicted abilities. Cross-validated Q^2 was defined as $Q^2 = (\mathrm{SSY} - \mathrm{PRESS})/\mathrm{SSY}$, where SSY was the sum of the squared deviations of the dependent variable values from their mean and PRESS was the predicted sum of squares obtained from the leave-one-out cross-validation method. The standard deviation of prediction (S_{PRESS}) was also considered and defined as $S_{\mathrm{PRESS}} = [\mathrm{PRESS}/(n - k - 1)]^{1/2}$, where k was the number of descriptors in the model and n was the number of compounds in the training date set. In addition, five compounds, selected from various ranges of anticonvulsant activity, were kept to test the actual prediction of the models.

logP: The hydrophobic coefficient of the molecules
π: The hydrophobic coefficient of the substitutes in sites 3, 4, 5
Area : The surface area of the molecules
Vm: The volume of the molecules
Hf : The final heat of formation of the molecules
MW: The molecular weight of the molecules
Density: The density of the molecule
$MR_{2,3,4}$: The total molar refraction in site 2, 3,4
Σσ: The hammett σ constant of the substituents on the benzene ring
Fh2o: The aqueous desolvation free energy of the molecules
Apol: Sum ot atomic polarizabilities of the molecules
homo, lumo: The energy of home and lumo orbitors of the molecules
Dip, Dip_x, Dip_y, Dip_z: The dipole vector and dipole vector components in x,y,z
Char_N: Atomic net charge of the O atom on the amido group
Char_O: Atomic net charge of the N atom on the amio group

Figure 1. Features used in the QSAR analysis of the data set.

RESULTS AND DISCUSSION

Construction of the Linear Polynomial QSAR Models. The data set contained 35 compounds and 19 molecular descriptors. The abbreviations for these descriptors are given in Figure 1. In our models, the five-term and six-term multiple linear regression models were constructed. More than six independent variables were not considered because of the rising possibility of chance correlation. For this data set, populations with 200 individuals were used, and the number of elite populations was defined as 100. The genetic operator was applied until the total fitness score of the elite populations no longer improved over a period of 30 evolution operations. Moreover, a partial reinitialization procedure was applied after 200 crossover operations. The convergence criterion was met after 1330 operations for 4 descriptors and 1760 operations for 5 descriptors.

After the calculations, the 100 best models for the 5 features and 4 features were obtained, respectively. The top 16 models selected from the 2 elite populations are listed in Table 3. Because a model could not be properly evaluated only by its multiple linear regression coefficient, the quality of the models, as indicated by SD, F, Q^2, and S_{PRESS}, was tested statistically. In eqs 1−16, n was the number of compounds used in the fit, SD being the standard error of mean, and F being the overall F statistics for the addition of each successive term, and the values in parentheses were the 95% confidence limit of each coefficient.

Generally, for the analysis of MLR, the data must be reduced to fewer and less correlated variables. The cross-correlated descriptors would mislead the QSAR model in uncovering the actual relationship between the biological activity and these descriptors. The correlation study of these descriptors in the top 16 models (see Table 2) had been performed, and eqs 1, 10, 12, 13, and 16 were all proven to contain 2 or more descriptors that were highly cross-correlated between each other.

To verify the models, a leave-one-out cross-validation procedure was carried out to the top 16 equations. Generally speaking, the leave-one-out cross-validation coefficient should be greater than 0.75. Considering this criterion, eqs 3, 5, 8, 9, and 11 were unsatisfactory, their predictive abilities being unacceptable. The statistical properties for the coefficients in eqs 2, 4, 6, 7, 11, and 15, i.e., the F values at the 0.95 confidence level, are summarized in Table 4. Equations 2, 7, 11, and 15 contained at least one insignificant coefficient, as revealed by the confidence interval and F statistics. After strict statistical verification, only models 4 and 6 were statistically significant, according to statistical criteria and predictive ability.

Table 2. Squared Correlation Matrix for Descriptors Appearing in The Best 16 QSAR Models Used in the Correlation Study

	Hf	Vm	Dip *z*	log *P*	$MR_{2,3,4}$	$\sum\sigma$	Fh2o	homo	lumo	density	area	dipole	π
Hf	1.000	−0.183	0.286	−0.102	0.295	−0.068	0.394	0.194	−0.147	0.144	−0.182	0.041	−0.045
Vm		1.000	−0.184	0.189	0.689	−0.236	0.010	0.344	0.054	−0.121	0.981	−0.031	0.199
Dip *z*			1.000	−0.156	−0.057	0.105	0.243	−0.048	−0.120	0.115	−0.204	−0.002	−0.073
log *P*				1.000	0.447	0.212	−0.219	−0.235	0.186	0.420	0.252	−0.063	0.955
$MR_{2,3,4}$					1.000	−0.237	−0.097	0.369	0.115	0.239	0.718	−0.143	0.519
$\sum\sigma$						1.000	0.513	−0.929	−0.789	0.404	−0.245	0.752	0.223
Fh2o							1.000	−0.481	−0.786	−0.027	−0.018	0.724	−0.099
homo								1.000	0.716	−0.379	0.331	−0.707	−0.299
lumo									1.000	−0.161	0.065	−0.792	0.059
density										1.000	−0.175	0.124	0.470
area											1.000	−0.052	0.255
dipole												1.000	−0.008
π													1.000

Table 3. Top 16 QSAR Models Generated by Using Training Data Set

1. $\log(1/C) = 1.212 + 0.323 \log P - 0.003$ Hf $- 0.007$Vm $+ 0.028$*Fh2o $- 0.014$ Dip *z*
   ($n = 30$, fitness = 0.862, SD = 0.621, $F = 13.870$, $Q^2 = 0.588$, $S_{PRESS} = 0.198$)
2. $\log(1/C) = -0.465 + 0.366 \log P - 0.157MR_{2,3,4} - 0.002$Hf $- 0.004$Vm $+ 0.023$Fh2o
   ($n = 30$, fitness = 0.862, SD = 0.754, $F = 13.857$, $Q^2 = 0.599$, $S_{PRESS} = 0.196$)
3. $\log(1/C) = 0.212 + 0.319 \log P - 0.003$area $- 0.011$Vm $- 0.003$Hf $+ 0.027$Fh2o
   ($n = 30$, fitness = 0.857, SD = 0.813, $F = 13.264$, $Q^2 = 0.499$, $S_{PRESS} = 0.219$)
4. $\log(1/C) = 1.125 + 0.327 \log P - 0.003$Hf $- 0.007$Vm $+ 0.027$Fh2o
   ($n = 30$, fitness = 0.856, SD = 0.821, $F = 17.095$, $Q^2 = 0.600$, $S_{PRESS} = 0.191$)
5. $\log(1/C) = 0.566 + 0.403 \log P - 0.330MR_{2,3,4} - 0.088$lumo $- 0.001$Hf $+ 0.011$Fh2o
   ($n = 30$, fitness = 0.855, SD = 0.624, $F = 13.024$, $Q^2 = 0.501$, $S_{PRESS} = 0.214$)
6. $\log(1/C) = 0.464 + 0.401 \log P - 0.327MR_{2,3,4} - 0.001$Hf $+ 0.017$Fh2o
   ($n = 30$, fitness = 0.853, SD = 0.689, $F = 16.614$, $Q^2 = 0.598$, $S_{PRESS} = 0.192$)
7. $\log(1/C) = 0.447 + 0.384MR_{2,3,4} + 0.043\pi + 0.032$Dip $- 0.001$Hf
   ($n = 30$, fitness = 0.852, SD = 0.913, $F = 16.580$, $Q^2 = 0.569$, $S_{PRESS} = 0.197$)
8. $\log(1/C) = -0.302 - 0.600MR_{2,3,4} + 0.479\pi + 0.003$area $- 0.016$Dip *z* $+ 0.129$density
   ($n = 30$, fitness = 0.852, SD = 0.724, $F = 12.709$, $Q^2 = 0.471$, $S_{PRESS} = 0.225$)
9. $\log(1/C) = -0.449 + 0.012\sum\sigma - 0.381MR_{2,3,4} + 0.429\pi + 0.030$Dip $- 0.001$Hf
   ($n = 30$, fitness = 0.852, SD = 0.834, $F = 12.735$, $Q^2 = 0.490$, $S_{PRESS} = 0.221$)
10. $\log(1/C) = -0.088 - 0.587MR_{2,3,4} + 0.487\pi + 0.003$area $- 0.016$Dip *z*
    ($n = 30$, fitness = 0.851, SD = 0.621, $F = 16.462$, $Q^2 = 0.543$, $S_{PRESS} = 0.205$)
11. $\log(1/C) = -1.693 - 0.539MR_{2,3,4} - 0.153$homo $+ 0.003$area $+ 0.002$Dip
    ($n = 30$, fitness = 0.851, SD = 0.754, $F = 12.571$, $Q^2 = 0.484$, $S_{PRESS} = 0.217$)
12. $\log(1/C) = -1.858 - 0.539MR_{2,3,4} + 0.443\pi - 0.170$homo $+ 0.003$area
    ($n = 30$, fitness = 0.850, SD = 0.921, $F = 16.349$, $Q^2 = 0.531$, $S_{PRESS} = 0.207$)
13. $\log(1/C) = -0.294 - 0.587MR_{2,3,4} + 0.488\pi + 0.024$Dip $+ 0.003$area
    ($n = 30$, fitness = 0.850, SD = 0.723, $F = 16.284$, $Q^2 = 0.529$, $S_{PRESS} = 0.208$)
14. $\log(1/C) = -0.688 - 0.341MR_{2,3,4} + 0.402 \log P - 0.190$lumo $- 0.001$Hf
    ($n = 30$, fitness = 0.850, SD = 0.763, $F = 16.259$, $Q^2 = 0.595$, $S_{PRESS} = 0.193$)
15. $\log(1/C) = -3.212 - 0.253MR_{2,3,4} - 0.001$Hf $+ 0.320 \log P - 0.305$homo
    ($n = 30$, fitness = 0.850, SD = 0.723, $F = 16.250$, $Q^2 = 0.596$, $S_{PRESS} = 0.193$)
16. $\log(1/C) = 0.520 + 0.149\sum\sigma - 0.354MR_{2,3,4} + 0.397\pi - 0.001$Hf
    ($n = 30$, fitness = 0.850, SD = 0.604, F = 16.015, $Q^2 = 0.548$, $S_{PRESS} = 0.203$)

To verify the actual prediction ability of these two models, five compounds which did not affect the model calibration were chosen as an external validation set. Table 4 shows the actual predictions of these two models for the five tested compounds. In terms of actual prediction, these two equations predicted well for the five tested compounds. The predictions were even better than those of the calibration set. For the value of SSE (see Table 5), eq 4 was more reliable than the others. Consequently, eq 4 could be considered to be the most suitable linear polynomial QSAR model to possess the best actual prediction ability. The predicted $\log(1/C)$ values for these 35 compounds are listed in Table 1.

Principal Features Determined. In most cases, the interactions between the molecular features were very complicated. The interaction between several features might result in another feature. Moreover, only from a single good model, we might not grasp the most original factors influencing biological activity. Information from multiple models might be more vital than that from a single model. So observation of the descriptor used in the multiple models could make it more reliable to uncover the actual vital descriptors. Figure 2 showed that the variables used changed along with the evolution procedure in the elite populations. The figure showed that after convergence, the frequency appearing in these models the elite population was quite different from that at the beginning. The frequencies of $MR_{2,3,4}$, π, log *P*, and Hf in the models were much higher than those of the other descriptors. The descriptor $MR_{2,3,4}$ was used in more than 80% of the models; π appeared in about 60% of the models. The next two descriptors, log *P* and Hf, respectively, were in about 55% and 40% of the models. They might be the most important factors affecting the biological activity. Besides these four factors, the frequencies of homo, $\sum\sigma$, lumo, and Dip also seemed relatively high. The appearance frequencies of the other 11 descriptors were very low in the elite population, so they might contribute little to the biological activity. From Table 3, it could be found that the top 8 features accounted for nearly all the features in the top 16 models. The values of these most important eight descriptors are listed in Table 6.

Table 4. The 95% Confidence, Level and *F* Statistics for the Coefficients of Variables in Eqs 1, 2, 4, 5, and 8−11

eq	variable	coeff	95% conf	*F*	significance
2	log *P*	0.323	±0.053	36.718	NS[a]
	Hf	−0.003	±0.001	24.852	
	Vm	−0.007	±0.002	18.438	
	Fh2o	0.028	±0.008	13.143	
	Dip *z*	−0.014	±0.014	0.993	
4	log *P*	0.327	±0.053	37.926	
	Hf	−0.003	±0.001	27.622	
	Vm	−0.007	±0.002	17.617	
	Fh2o	0.027	±0.008	12.388	
6	log *P*	0.401	±0.060	44.687	
	$MR_{2,3,4}$	−0.327	±0.080	16.740	
	Hf	−0.001	±0.001	5.555	
	Fh2o	0.017	±0.008	4.964	
7	$MR_{2,3,4}$	−0.384	±0.084	21.161	NS
	π	0.432	±0.066	42.803	NS
	Dip	0.032	±0.022	2.065	
	Hf	−0.001	±0.001	3.639	
11	log *P*	0.402	±0.060	44.247	NS
	$MR_{2,3,4}$	−0.341	±0.080	18.367	
	lumo	−0.190	±0.089	4.499	
	Hf	−0.001	±0.001	3.519	
15	log *P*	0.320	±0.068	21.918	NS
	$MR_{2,3,4}$	−0.253	±0.093	7.390	
	Hf	−0.001	±0.001	2.724	
	homo	−0.305	±0.144	4.484	

[a] Not significant.

Table 5. Actual Prediction (Eqs 4, 6, and 18) for the Five Compounds

		eq 4		eq 6		eq 18	
compd	log(1/*C*) expt	pred	residue	pred	residue	pred	residue
6	0.550	0.647	0.097	0.694	0.154	0.621	0.071
10	0.921	0.815	0.106	0.851	0.070	0.899	0.022
17	0.290	0.332	0.042	0.350	0.060	0.348	0.058
23	0.735	0.732	0.003	0.759	0.024	0.684	0.051
28	0.620	0.570	0.050	0.488	0.132	0.600	0.020
SSE[a]			0.025		0.050		0.012

[a] Sum of squares error of prediction for five tested compounds.

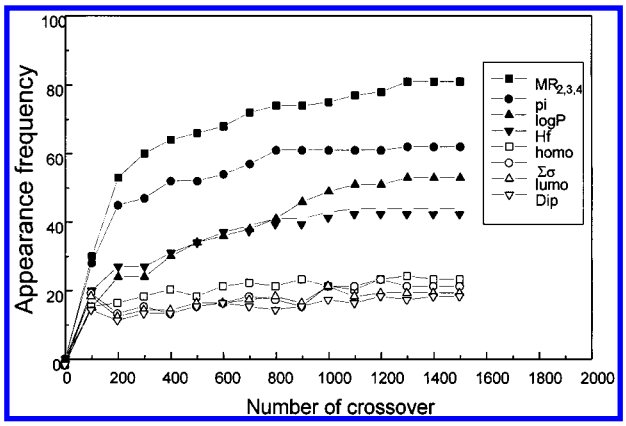


Figure 2. Change in the descriptor used in the evolution procedure of the elite population with four descriptors.

Due to the high appearance frequencies of $MR_{2,3,4}$, π, log *P*, and Hf, it seemed that these four variables played strong roles in the proposed QSAR models. According to the $MR_{2,3,4}$ definition, which stood for molar refraction of subsitituents in sites 2, 3, and 4 on the benzene ring, the negative coefficient of it pointed out that small groups on the benzene ring contributed to the high biological activity. It was suggested from Figure 2 that π was a necessary contributor

Table 6. Top Eight Features Derived from Figure 2

no.	π	$MR_{2,3,4}$	log *P*	Hf	homo	$\Sigma\sigma$	lumo	Dip
1	0.71	0.80	3.43	−4.329	−9.423	0.37	−0.577	4.112
2	0.14	0.29	2.86	−42.641	−9.422	0.34	−0.597	4.263
3	0.14	0.29	2.86	−43.064	−9.305	0.06	−0.554	2.864
4	0.86	1.09	3.57	6.930	−9.364	0.23	−0.656	2.873
5	1.42	1.30	4.14	−7.658	−9.448	0.46	−0.819	3.940
6	1.42	1.30	4.14	−7.047	−9.475	0.60	−0.840	4.160
7	0.71	0.80	3.43	−4.543	−9.347	0.23	−0.603	2.855
8	−0.88	0.94	2.44	28.839	−9.763	0.78	−1.875	6.285
9	−0.28	0.94	2.44	29.402	−9.705	0.71	−1.726	8.423
10	0.88	0.70	3.60	−148.750	−9.522	0.43	−0.726	5.119
11	0.88	0.70	3.60	−145.368	−9.485	0.54	−0.735	5.494
12	0.88	0.70	3.60	−149.105	−9.558	0.54	−0.884	3.292
13	−0.69	1.17	2.05	−70.834	−8.772	−0.15	−0.383	1.642
14	−0.02	0.99	2.70	−33.821	−8.923	−0.27	−0.287	2.314
15	1.12	1.60	3.84	−19.837	−9.435	0.35	−0.576	4.095
16	0.47	1.45	3.19	−40.318	−8.896	−0.24	−0.274	2.393
17	1.05	1.91	3.77	−46.351	−8.892	−0.25	−0.270	2.487
18	1.55	2.37	4.27	−52.948	−8.894	−0.32	−0.266	2.572
19	0.71	0.80	3.07	−10.960	−9.473	0.37	−0.565	4.334
20	0.14	0.29	3.10	−49.429	−9.472	0.34	−0.584	4.431
21	0.14	0.29	3.10	−49.784	−9.311	0.06	−0.563	3.081
22	0.86	1.09	3.82	0.187	−9.379	0.23	−0.652	3.120
23	1.42	1.30	4.38	−14.922	−9.502	0.46	−0.794	4.052
24	1.42	1.30	4.38	−14.324	−9.516	0.60	−0.823	4.301
25	0.71	0.80	3.67	−11.238	−9.361	0.23	−0.603	3.100
26	0.88	0.70	3.84	−156.297	−9.743	0.54	−0.865	3.459
27	0.88	0.70	3.84	−155.987	−9.660	0.43	−0.713	5.364
28	0.71	0.80	3.33	−4.952	−9.480	0.37	−0.569	4.312
29	0.14	0.29	2.76	−43.834	−9.314	0.06	−0.564	3.116
30	0.86	1.09	3.84	6.152	−9.383	0.23	−0.655	3.155
31	1.42	1.30	4.04	−9.021	−9.507	0.46	−0.788	4.027
32	0.71	0.80	3.33	−5.262	−9.364	0.23	−0.605	3.135
33	1.42	1.30	4.04	−8.339	−9.523	0.60	−0.825	4.263
34	0.88	0.70	3.50	−150.343	−9.747	0.54	−0.875	3.576
35	0.88	0.70	3.50	−150.086	−9.678	0.43	−0.714	5.111

to the anticonvulsant activity, and this variable represented the partition coefficient affected by the substituents on the benzene ring. A positive sign of the coefficient for this term indicated that high hydrophobic substituents on the benzene ring were very vital to the anticonvulsant activity. So it could be reasonably presumed that the benzene ring combined with these substituents on it was composed of a large hydrophobic core. This hydrophobic group would produce a strong hydrophobic interaction with the receptor. The negative sign of $MR_{2,3,4}$ indicated that when benzene ring interacted with the receptor through hydrophobic interaction, the steric space may be relatively small and the existence of the substituents on the benzene ring would hinder the most adequate orientation of the inhibitor and the receptor in order to produce the best hydrophobic interaction. The anticonvulsant activity could be largely explained by these two descriptors, $MR_{2,3,4}$ and π:

$$\log(1/C) = 0.659 - 0.461MR_{2,3,4} + 0.470\pi$$
$$(n = 30,\ r = 0.819,\ F = 27.357,\ \mathrm{SD} = 0.651) \quad (17)$$

The parameter log *P* seemed also very important to the biological activity. But the correlation study showed that log *P* was not an independent feature, which was highly cross-correlated with π, and the correlation coefficient was 0.955. That is to say, the change of the value of log *P* was mainly caused by the change of the partition coefficient of the substituents on the benzene ring. From the above analyses, the property of the substituents on the benzene ring was critical to the biological activity.

From Figure 2, the formation energy of the molecules contributed a lot to the biological activity. From Tables 1

and 6, it could be found that the values of this parameter were mainly affected by the element constitution of substituents on the benzene ring. For example, for three different substitutes with F, Cl, or Br atoms respectively, the molecules with Br atoms on the benzene ring possessed a relatively smaller formation energy than the other molecules with Cl or F atoms on the benzene ring. This parameter was determined by the intrinsic properties of the molecules.

Compared with these four parameters, several other descriptors with relatively high frequencies contributed a little to the value of log(1/*C*). They were not very well-distinguished from other parameters in usage. Addition of these descriptors to eq 17 would cause an improvement of the *r* value, but the improvement was not very significant. From the correlation study, it could be found that the homo, lumo, and Dip were highly cross-correlated with $\sum\sigma$, and the changes of the FMO energy and the dipole were affected by the electronic properties of the substituents on the benzene ring. It could be presumed that the effect of the anticonvulsant activity of $\Sigma\sigma$ would influence the polarization of the amido carbonyl group through the conjugated effect. High $\sum\sigma$ would produce a high dipole and, consequently, and would enhance the dipole−dipole interaction between the ligands and the receptor. So we could conclude that the electronic effect would influence the anticonvulsant activity, but the contribution was relatively small.

Construction of the Linear Spline QSAR Models. To inspect these important factors more deeply, several descriptors were chosen and linear spline models were constructed. Through construction of spline models, it was expected that we would discover whether these features were predictive only in a limited range of values or not. Four molecular features, $MR_{2,3,4}$, π, Hf, and $\sum\sigma$, were selected for constructing linear spline models. The correlation study had shown that these four features were not cross-correlated to each other, which meant they were all independent features and had their own independent contributions to the anticonvulsant activity. The splines used here were truncated power splines and were denoted with angled brackets. For example, $\langle f(x) - a\rangle$ was equal to zero if the value of $f(x) - a$ was negative; otherwise, it was equal to $f(x) - a$. The regression with splines allowed the incorporation of features that did not have a linear effect over their entire range. But it was well-known that when we constructed the spline models, if the variables selected were truly linear in their biological activity, splines would not discover any more-predictive models but might confuse the model building with a chance correlation. So these models using spline terms must be carefully verified in order to test their validities.

The five-term models were constructed and evaluated with their regression coefficient as the fitness score. QSAR analysis began with a population of 200 random models. The population was converged after 850 crossover operations. The best model gained from the elite population is

$$\log(1/C) = 0.899 - 0.823\langle 0.70 - MR_{2,3,4}\rangle - 0.008\langle Hf + 40.318\rangle - 1.147\langle 0.23 - \sum\sigma\rangle - 1.792\langle -0.28 - \pi\rangle \quad (n = 30,\ r^2 = 0.820,\ F = 28.800,\ Q^2 = 0.744,\ S_{PRESS} = 0.154) \quad (18)$$

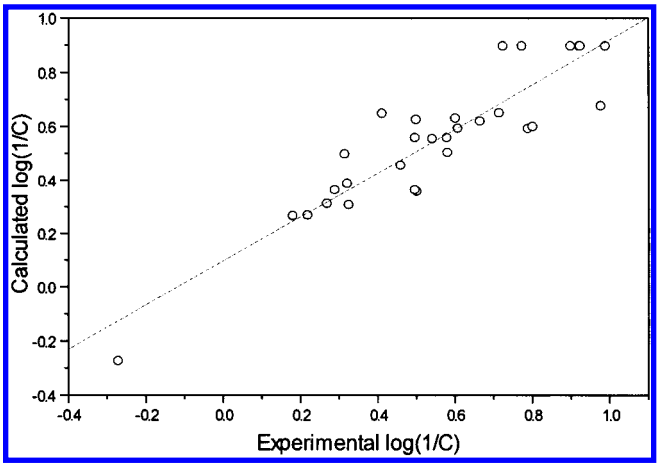


Figure 3. Comparison of experimental log(1/*C*) with calculated log(1/*C*) obtained from eq 18.

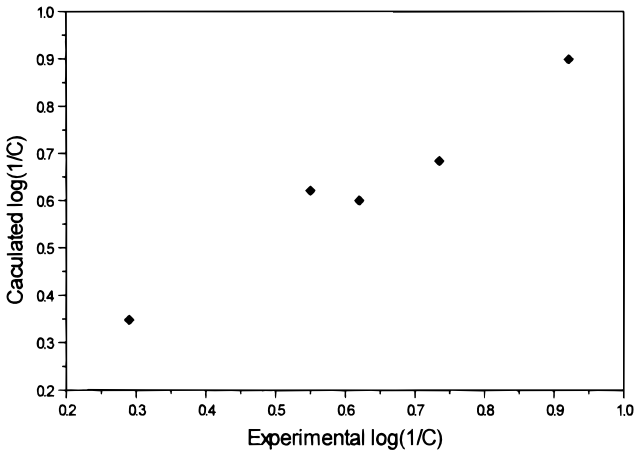


Figure 4. Plot of the actual prediction of eq 18 for five tested compounds.

The statistical result of eq 18 showed that this spline model seemed much better than the linear regression models in Table 2. The values of *r*, *F*, Q^2, and S_{PRESS} were all improved to some extent, compared with those in eqs 1−16. The actual prediction of this model was verified, the five tested compounds were predicted by eq 18 (see Table 5 and Figures 3 and 4), and the sum of squares error of prediction was much smaller than that of eq 4. The high statistical significance and high predictive ability showed that eq 18 was an excellent model.

From this model, the knot of every parameter could be obtained and could tell us the information about the range identifications for these four features. Hf would produce a negative contribution to the anticonvulsant activity when the value of Hf is lower than 40.318. A high $\sum\sigma$ was preferred when it was not higher than 0.23. From Table 6, it could be found that there were only eight compounds whose electronic effect of the substituents on the benzene ring would affect the anticonvulsant activity. A high value of π brought on high anticonvulsant activity, but when π was greater than 0.28, the influence would no longer increase with the increments of the π values of the substituents on the benzene ring. When the value of π was greater than 0.28, the steric effect and the electronic effect would influence the anticonvulsant activity mainly. A high value of $MR_{2,3,4}$ was preferred, as long as the value was below 0.7. An increment of the molar refraction of the substituents on the benzene

ring was favorable to anticonvulsant activity. This conclusion was different from that of the linear regression models. Linear regression models showed that the small substituents on the benzene ring were more favorable. It was not strange because the influence of $MR_{2,3,4}$ was not linearly effective during all ranges, and when it was treated as a linear polynomial term, the result could be quite different. When the inhibitor interacted with its receptor, the steric complementarity was expected to be optimal. When the volume of some parts of the inhibitor increased, the contact area between the inhibitor and the receptor might become larger. However, when the contact area increased to a certain value, the steric complementarity would not improve and might even be depressed by the steric hindrance. So there should exist an optimal value of $MR_{2,3,4}$; the value below or above it might produce the different contributions to the anticonvulsant activity. Both the linear regression model and the linear spline model were correct in this issue to some extent, but they expressed different features of this parameter.

The usage of splines must be careful. If the variables selected were truly linear in their effect on the biological activity, splines would not discover any more-predictive models and might confuse the model building with chance correlations. For eq 18, based on statistical criteria and the actual predictive ability for both internal and external sets of compounds, it had been proven that this model was a suitable correlation equation. So the results from the linear spline model may uncover the underlying mechanism of activity and express the actual relationship between the anticonvulsant activity and the molecular descriptors.

CONCLUSIONS

In this study, we attempted to correlate the antisultant activity with a lot of molecular properties. By using a GA, the linear regression models were constructed. These derived models were acceptable from the viewpoint of statistical significance and actual predictive ability. From the analyses of the descriptors used during the evolution procedure, the principal features relevant to the biological activity were identified. Through constructing the linear spline models, the effective ranges were determined for these four principal components, including the partition coefficient, molar refraction, Hammett σ constant of the substituent on the benzene ring, and the formation energy of the molecules.

ACKNOWLEDGMENT

We are particularly grateful to Prof. R. L. Li of Beijing Medical College for his excellent work on the synthesis and bioactivity test for these compounds used in our study. This work was supported by the NCSF 29992590-2 and 29573095 of P. R. China. Some source codes used in this study can be obtained from us upon request.

REFERENCES AND NOTES

(1) Pi, Y. Q. *J. Chin. Med.* **1978**, *58*, 216−211.
(2) Zhang, X. H.; Li, R. L.; Cai, M. S. *Beijing Med. College Trans.* **1980**, *12*, 83−91.
(3) Moffet, R. B. *J. Med. Chem.* **1964**, *7*, 319−325.
(4) Van Heyningen, E.; Brown, C. N. *J. Med. Chem.* **1966**, *6*, 675−681.
(5) Wang, Y. S.; Li, R. L.; Liu, W. Q.; Wang, G. Q.; Liu, P.; Xong, J. M.; Pei, Y. Q.; Yao, H. Y.; Gao, X. M. *Acta Pharm. Sin.* **1986**, *21*, 542−545.
(6) Wang, Y. S.; Li, R. L.; Liu, W. Q.; Xu, X. J.; Guan, Y. *Org. Chem.* **1986**, *8*, 217−220.
(7) Hou, T. J.; Wang, J. M.; Xu, X. J. *Chemom. Intell. Lab. Syst.*, in press.
(8) Li, R. L.; Wang, Y. S. *Acta Pharm.* **1986**, *21*, 580−585.
(9) InsightII, User's Guide, Molecular Simulation Inc., 1997.
(10) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. *Proteins: Struct., Funct., Genet.* **1988**, *4*, 31−47.
(11) Mopac 7.0, User's Guide, Quantum Chemistry Program Exchange, 1993.
(12) Hopfinder, A. J. *Conformational Properties of Macromolecules*; Academic Press: New York, 1977.
(13) Roger, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(14) Leardi, R.; Boggia, R.; Terrile, M. *J. Chem.* **1992**, *6*, 267−281.
(15) Leardi, R. *J. Chem.* **1994**, *8*, 67−79.

CI990010N