Genetic Algorithm Guided Selection: Variable Selection and Subset Selection

Sung Jin Cho*,† and Mark A. Hermsmeier§

New Leads, Bristol-Myers Squibb Co., 5 Research Parkway, Wallingford, Connecticut 06492-7660, and
New Leads, Bristol-Myers Squibb Co., P.O. Box 4000, Princeton, New Jersey 08543-4000

Received November 30, 2001

A novel Genetic Algorithm guided Selection method, GAS, has been described. The method utilizes a simple encoding scheme which can represent both compounds and variables used to construct a QSAR/QSPR model. A genetic algorithm is then utilized to simultaneously optimize the encoded variables that include both descriptors and compound subsets. The GAS method generates multiple models each applying to a subset of the compounds. Typically the subsets represent clusters with different chemotypes. Also a procedure based on molecular similarity is presented to determine which model should be applied to a given test set compound. The variable selection method implemented in GAS has been tested and compared using the Selwood data set ($n = 31$ compounds; $v = 53$ descriptors). The results showed that the method is comparable to other published methods. The subset selection method implemented in GAS has been first tested using an artificial data set ($n = 100$ points; $v = 1$ descriptor) to examine its ability to subset data points and second applied to analyze the XLOGP data set ($n = 1831$ compounds; $v = 126$ descriptors). The method is able to correctly identify artificial data points belonging to various subsets. The analysis of the XLOGP data set shows that the subset selection method can be useful in improving a QSAR/QSPR model when the variable selection method fails.

INTRODUCTION

The idea of quantitative structure−activity (or structure−property) relationships (QSAR/QSPR) was introduced by Hansch et al. in 1963 and first applied to analyze the importance of lipophilicity for biological potency.[1−3] This concept is based on the assumption that the difference in the structural properties of molecules, whether experimentally measured or calculated, accounts for the difference in their observed biological or chemical properties. In traditional medicinal chemistry, QSAR/QSPR models are obtained from a training set typically consisting of structurally similar analogues. The models attempt to identify important structural features of molecules that are relevant to explaining variations in biological or chemical properties. The problem of feature selection has been the focus of many research efforts.[4−13] As the structural diversity in a QSAR/QSPR training set increases, constructing a good model becomes increasingly difficult, and simply performing variable selection might not improve the quality of the model. In this case, the quality of the model may depend less on the types of variables present and more on the types of compounds present in the data set. Our approach attempts to extend the applicability of QSAR/QSPR to more structurally diverse data sets. Good examples of such data sets include percent inhibition or IC_{50} values obtained from high throughput screening, cell-based assays, assays with multiple mechanisms of action, or a data set created by pooling absorption, distribution, metabolism, or excretion (ADME) data from various therapeutic programs. It also provides an unbiased mechanism for removing outliers from a model.

In this paper, we report the genetic algorithm[14−16] guided selection (GAS) method. This method utilizes a simple encoding scheme that can represent both compounds and descriptor variables used to construct a QSAR/QSPR model. The GA optimization can then be used to select either variables (variable selection) or compounds (subset selection) depending on the problem. The variable selection implemented in the method is tested using the well-known Selwood data set[4] and compared with other published results.[5−10] The subset selection concept is first tested using an artificial data set to examine the method's ability to subset data points. The method is then used to develop a QSPR model for the XLOGP data set.[17] In this case the newly developed binding interaction pair fingerprint (BIPF) descriptors are used as the variables. The result is again compared with other published results.[17,18,19] The general utility of the subset selection method in the construction of a QSAR/QSPR model is discussed.

DATA SETS AND DESCRIPTORS

Selwood Data Set. The Selwood data set contains 31 Antifilarial Antimycin A_1 analogues tested for in vitro Antifilarial activity (Table 1). Analogues were described using 53 different physicochemical parameters from MolconnZ. The descriptions of variables used are listed in Table 2. Both biological activities and descriptors are used as originally reported in ref 4 in order to compare with the published results (Table 5).

Parabolic Data Set. An artificial data set is created by using the following simple equation of a parabola:

$$\text{Activity} = \text{Descriptor}^2 \qquad (1)$$

The descriptor for this artificial data set ranges from −50 to 49, making the number of data points to be 100.

* Corresponding author phone: (805)447-4269; fax: (805)480-3015; e-mail: scho@amgen.com.
† New Leads, Wallingford, CT.
§ New Leads, Princeton, NJ.

10.1021/ci010247v CCC: $22.00
Published on Web 05/23/2002

Table 1. -log in Vitro Activity (EC50 in μM) of the Selwood Data Set[4]

compd no.	-log EC50	compd no.	-log EC50	compd no.	-log EC50
1	−0.85	11	0.10	21	0.82
2	−0.38	12	1.13	22	1.36
3	1.40	13	0.92	23	0.23
4	0.32	14	0.77	24	1.41
5	−0.88	15	0.30	25	−0.04
6	0.82	16	1.36	26	0.43
7	1.84	17	−1.00	27	1.03
8	1.02	18	−0.41	28	1.55
9	0.42	19	−0.90	29	1.07
10	0.00	20	0.89	30	−1.00
				31	0.48

Table 2. Variables Used in the Selwood Data Set[4]

variable	description
ATCH1	partial atomic charge for atoms 1
ATCH2	partial atomic charge for atoms 2
ATCH3	partial atomic charge for atoms 3
ATCH4	partial atomic charge for atoms 4
ATCH5	partial atomic charge for atoms 5
ATCH6	partial atomic charge for atoms 6
ATCH7	partial atomic charge for atoms 7
ATCH8	partial atomic charge for atoms 8
ATCH9	partial atomic charge for atoms 9
ATCH10	partial atomic charge for atoms 10
ESDL1	electrophilic superdelocalizability for atom 1
ESDL2	electrophilic superdelocalizability for atom 2
ESDL3	electrophilic superdelocalizability for atom 3
ESDL4	electrophilic superdelocalizability for atom 4
ESDL5	electrophilic superdelocalizability for atom 5
ESDL6	electrophilic superdelocalizability for atom 6
ESDL7	electrophilic superdelocalizability for atom 7
ESDL8	electrophilic superdelocalizability for atom 8
ESDL9	electrophilic superdelocalizability for atom 9
ESDL10	electrophilic superdelocalizability for atom 10
NSDL1	nucleophilic superdelocalizability for atom 1
NSDL2	nucleophilic superdelocalizability for atom 2
NSDL3	nucleophilic superdelocalizability for atom 3
NSDL4	nucleophilic superdelocalizability for atom 4
NSDL5	nucleophilic superdelocalizability for atom 5
NSDL6	nucleophilic superdelocalizability for atom 6
NSDL7	nucleophilic superdelocalizability for atom 7
NSDL8	nucleophilic superdelocalizability for atom 8
NSDL9	nucleophilic superdelocalizability for atom 9
NSDL10	nucleophilic superdelocalizability for atom 10
DIPV_X	dipole vector in X direction
DIPV_Y	dipole vector in Y direction
DIPV_Z	dipole vector in Z direction
DIPMOM	dipole moment
VDWVOL	van der Waals volume
SURF_A	surface area
MOFI_X	principal moments of inertia in the X direction
MOFI_Y	principal moments of inertia in the Y direction
MOFI_Z	principal moments of inertia in the Z direction
PEAX_X	principal ellipsoid axes in the X direction
PEAX_Y	principal ellipsoid axes in the Y direction
PEAX_Z	principal ellipsoid axes in the Z direction
MOL_WT	molecular weight
S8_1DX	substituent dimensions in the X direction
S8_1DY	substituent dimensions in the Y direction
S8_1DZ	substituent dimensions in the Z direction
S8_1CX	substituent centers in the X direction
S8_1CY	substituent centers in the Y direction
S8_1CZ	substituent centers in the Z direction
LOGP	partition coefficient
M_PNT	melting point
SUM_F	sum of F substituent constant
SUM_R	sum of R substituent constant

Table 3. Top 10 Models Obtained after the Variable Selection Analysis of the Selwood Data Set

parent	variable	r^2	s	F	Q^2 [a]	SDEP[b]
1	3	0.721	0.460	23.316	0.647	0.483
2	3	0.720	0.461	23.094	0.645	0.485
3	3	0.719	0.462	22.986	0.644	0.485
4	2	0.610	0.534	21.905	0.534	0.555
5	3	0.702	0.476	21.175	0.605	0.511
6	3	0.697	0.479	20.735	0.601	0.514
7	2	0.602	0.540	21.194	0.524	0.561
8	2	0.599	0.542	20.872	0.518	0.564
9	2	0.597	0.543	20.746	0.501	0.574
10	3	0.689	0.486	19.910	0.599	0.515

[a] Leave one out cross-validation. [b] Standard Error of Prediction.

XLOGP Data Set. The XLOGP data set contains 1831 compounds with their experimental log *P* values, which were used to develop an atom-additive log P model.[17] The constructed XLOGP model was then used to predict the log *P* values of 19 drugs which were not in the training set.[18] The key step in building the XLOGP model in the reference implementation is the classification of each atom, which is based on its hybridization state and neighboring atoms. Contributions associated with 80 different atom classes and five correction factors were then obtained by performing multiple linear regression (MLR) analysis on 1831 compounds. The log *P* value of a compound is estimated by identifying existing atom classes and correction factors and summing up their contributions. Rather than using reported 80 atom classes and five correction factors, we have developed and used binding interaction pair fingerprint descriptors instead. The BIPF descriptors are derived from the definition of physicochemical property descriptors developed by Kearsley et al.,[20] which is a variant of atom pair descriptors.[21] The atom pair descriptors are defined as all possible unique atom pairs with the shortest bond path and their occurrence. Rather than using atoms to generate pairs, Kearsely et al. extended the definition by creating property classes (such as binding property, charge, and hydrophobic classes) to replace atom classes. The idea behind this approach is to perceive physicochemically equivalent atoms and classify them as one group. In BIPF descriptors, atoms are grouped into any of six binding interaction classes: H-donor, H-acceptor, acid, base, hydrophobe, and aromatic ring classes. The bond path information is also simplified by utilizing six bond path bins (Bin 1: 1 & 2; Bin 2: 3 & 4; Bin 3: 5 & 6; Bin 4: 7 & 8; Bin 5: 9 & 10; Bin 6: 11 & up). The fingerprint is then simply the combination of two binding interaction pairs (out of 36 only 21 is unique) and their shortest bond path bins (6); the count information is omitted. The total number of possible combinations is thus 126 and represents the size of the fingerprint. The key difference between BIPF and original XLOGP descriptors is that BIPF does not rely on the importance of element type to define an atom class but emphasizes the potential for binding interactions to define the atoms. The internally developed program SMI2BIPF (converts SMILES to BIPF) is written in C and uses the Daylight SMARTS toolkit for atom class perception.[22] Compounds used as the training (n = 1831) and test (n = 19) sets in the development of the XLOGP model[17] were converted to BIPF descriptors and used in our analyses.

Table 4. Top 10 QSAR Models Obtained after the Variable Selection Analysis of the Selwood Data Set

parent	equation
1	Act = −2.503 − 0.000075 * MOFI_Y + 0.584 * LOGP + 1.513 * SUM_F
2	Act = 2.872 + 0.809 * ESDL3 − 0.0130 * SURF_A + 0.569 * LOGP
3	Act = −2.537 − 0.000073 * MOFI_Z + 0.569 * LOGP + 1.452 * SUM_F
4	Act = −2.219 − 0.000083 * MOFI_Z + 0.679 * LOGP
5	Act = −0.806 + 0.735 * ESDL3 − 0.000077 * MOFI_Y + 0.589 * LOGP
6	Act = −0.932 + 0.692 * ESDL3 − 0.000075 * MOFI_Z + 0.574 * LOGP
7	Act = −2.151 − 0.000084 * MOFI_Y + 0.694 * LOGP
8	Act = −0.227 − 0.207 * PEAX_X + 0.609 * LOGP
9	Act = 1.647 − 0.0138 * SURF_A + 0.667 * LOGP
10	Act = −0.777 − 0.177 * PEAX_X + 0.504 * LOGP + 1.343 * SUM_F

Table 5. Comparison of Variables Selected Using Different Methods

	methods						
variable	Selwood[4]	Wikel[5]	McFarland[6]	Rogers[7]	Kubinyi[9]	Waller[10]	GAS[a]
ATCH1				X	X	X	X
ATCH2	X	X	X				
ATCH3					X	X	X
ATCH4		X	X	X	X	X	X
ATCH5			X	X	X		X
ATCH6				X	X		
ATCH7					X	X	X
DIPV_X		X	X		X		
DIPV_Y	X						
DIPV_Z	X						
ESDL3				X	X	X	X
ESDL5	X						
ESDL8						X	
ESDL10	X						
NSDL2	X						
VDWVOL		X	X		X	X	X
SURF_A				X	X	X	X
MOFI_X		X	X			X	
MOFI_Y		X	X	X	X		X
MOFI_Z					X	X	X
PEAX_X				X	X	X	X
PEAX_Y		X	X				
S8_1DX			X				
S8_1CZ	X						
LOGP	X	X	X	X	X	X	X
M_PNT	X	X	X		X		X
SUM_F				X	X	X	X
SUM_R	X						

[a] Top 25 models.

METHOD

Genetic Algorithm Guided Selection (GAS). As with any genetic algorithm based optimization method, encoding scheme and fitting function represent two major steps in the GAS method. The encoding scheme implemented in GAS is shown in Figure 1. A typical QSAR table consists of compound rows and descriptor columns. The optimization goal is to find a combination of descriptors (variable selection to generate a set of descriptors) or a combination of compounds (subset selection to group compounds). The encoding scheme employed should be able to capture these combinations in a linear format in order to apply the GA operations, crossovers, and mutations. To represent such combinations, each parent (or combination) in a population contains indicator variables to control the presence or absence of compounds and variables (Figure 1). For the variable selection method, all indicator variables representing compounds are turned on (set to 1), but those representing descriptors are either turned on or off (set to 1 or 0 respectively) according to the presence or absence of variables. For the subset selection method the indicator variables representing compounds designate the subset that each compound is provisionally assigned to. Figure 1 shows the three subset (or equation) models and illustrates that each compound belongs to one of three subsets. The fitness of each parent is calculated using the following equation if the denominator term, (n − CVR × v), is greater than 1

$$\text{Fitness} = 1 - n * \sum(\text{Act} - \text{Cal})^2/(n - \text{CVR} \times v) \quad (2)$$

where n is the number of compounds, $\sum(\text{Act} - \text{Cal})^2$ is the sum of squared differences between actual and calculated activities of every compounds in a given set, CVR is the compound-to-variable ratio factor, and v is the number of variables. The fitness of a parent is severely penalized if the denominator term is not greater than 1, ensuring the model to maintain a specific compound-to-variable ratio as defined using the CVR parameter, thus minimizing overfitting. For the subset selection method, the fitness of each parent is simply the average fitness over all subset models. The subset selection encoding scheme shown in Figure 1, for example, would produce three models per parent, and the fitness of each parent is the average fitness of three models.

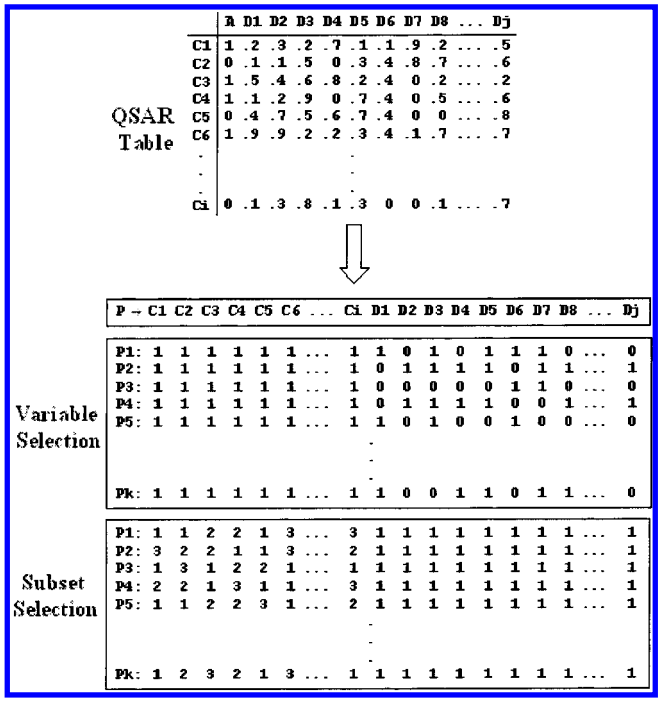

Figure 1. Encoding scheme used in genetic algorithm guided selection (i − number of compounds; j − number of descriptors; k − number of parents).

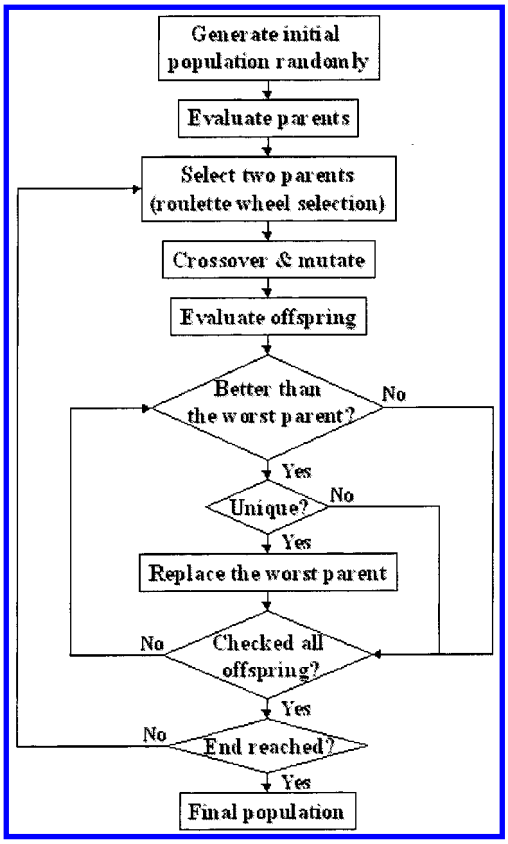

Figure 2. Genetic algorithm guided selection.

The overall steps involved in the GAS method is illustrated in Figure 2. The process begins with the random generation of an initial population containing a specified number of parents (encoding scheme shown in Figure 1). After the fitness of each parent is evaluated using eq 2, two parents are picked based on the roulette wheel selection[14] method; a high fitting parent has a higher probability of being selected. Offspring are generated by performing a crossover or a mutation. In a crossover, part of the genetic makeup of one parent is mixed with part of the genetic makeup of the other parent at a randomly selected crossover point. In a mutation, an individual gene, again selected at random, is changed to either one or zero for the variable selection and one compound is selected at random to move from one subset to another for the subset selection. Crossover and mutation points are appropriately limited to those regions where variation is allowed; the regions encoding compounds and variables are conserved for the variable selection and the subset selection, respectively. The fitness of each offspring is calculated and compared with the fitness of parents in the population. If the fitness of an offspring is better than the lowest scoring parent, the offspring replaces the worst parent. Otherwise, two new parents are selected again from the population, and the entire process is repeated for a specified number of times.

Prediction. The prediction of compounds in the test set using the models obtained via the subset selection method presents a special challenge. To remedy the problem of which model is used to predict which test set compounds, the test set compounds are assigned to one of the subsets based on their similarities (defined via Euclidean distance calculated using selected descriptors) to compounds in the training set. If compound A in the test set is similar to compound B in the training set, the model which includes B is used to predict A. The output of the prediction consists of the subset number in which the test compound is belong to, a compound in the training set that is most similar to the test compound, their similarity, and the predicted activity using the model in which the training set compound belongs to.

Predictive r^2. A predictive test is applied to determine the optimal number of subsets. For the subset selection method, a predictive r^2 value is calculated to estimate the ability of one model to predict the compounds belonging to other models, when all models have high r^2 values; if they do not have high r^2 values, calculating the predictive r^2 values is not necessary. If the predictive r^2 values of two models, generated by predicting against each other, are similar, the subset selection might not be necessary for compounds in these models because one of the models might be sufficient in describing compounds in both models. If they are not similar, the subset selection method is necessary to better correlate the data set. The predictive r^2 is defined as

$$\text{Predictive } r^2 = 1 - \sum(\text{Act} - \text{Prd})^2 / \sum(\text{Act} - \text{Avg})^2 \quad (3)$$

where $\sum(\text{Act} - \text{Prd})^2$ is the sum of the squared differences between the actual and predicted activities of every compounds in the test set, and $\sum(\text{Act} - \text{Avg})^2$ is the sum of squared differences between the actual activities of compounds in the test set and the average activity of the training set compounds.[23]

Implementation. SMI2BIPF and GAS programs have been implemented as C programs on a R10000 SGI server. The number of GA optimization steps and the size of a population are set to 10 000 and 100, respectively, for all GAS runs except otherwise noted. The CVR factor is set to 5 to maintain at least five compounds per variable ratio, thus minimizing overfitting.

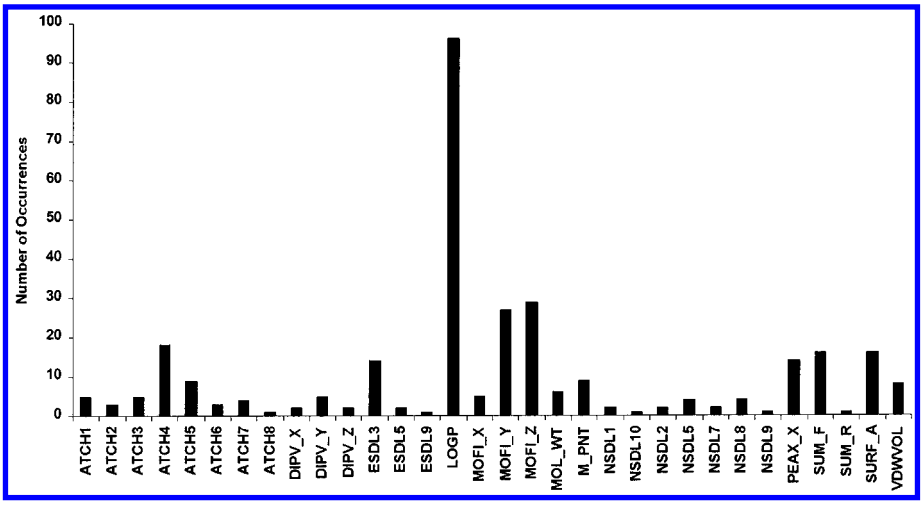

Figure 3. Number of occurrences of variables used in 100 parents in the final population selected after the variable selection of the Selwood data set.

RESULTS

Selwood Data Set. Variable Selection was applied to the Selwood data set. Tables 3 and 4 show the result of this analysis, with the statistics of the 10 best models listed in Table 3. Table 4 shows the corresponding QSAR equations. For these 10 models the number of variables range from 2 to 4, r^2 ranges from 0.517 to 0.777, and F ranges from 14.815 to 23.315. The best 3 variable model identified (Parent 1) consists of MOFI_Y, LOGP, and SUM_F descriptors and is identical to the best 3 variable model identified by Rogers and Hopfinger[7] and Kubinyi.[9] The third best 3 variable model identified (Parent 3), which consists of MOFI_Z, LOGP, and SUM_F, was not identified by Rogers and Hopfinger[7] but later identified by Kubinyi,[9] and it is encouraging to find this model on the top of the list. For the entire population of 100 unique solutions there are 32 unique descriptors used in various combinations. The number of occurrences of these descriptors is shown in Figure 3. The LOGP descriptor was used most frequently (96 out of 100 models), followed by MOFI_Z and MOFI_Y. Table 5 shows the comparison of the variables selected using different methods. The last column (GAS) represents the descriptors found in the top 25 models. The GAS method identified most of interesting descriptors and was found to be very comparable to other selection methods shown in Table 5. The entire GAS run took ~8 s in a single processor R10000 SGI server.

Subset Selection of the Parabolic Data Set. The subset selection method was first tested using an artificial data set to examine the method's ability to subset data points. The parabolic data set containing 100 points was generated and used as the artificial training set. The number of GA optimization steps was set to 20 000 for all GAS analyses of the parabolic data set. The selection of this training set, containing nonlinear data structure, is an attempt to mimic the complex nature of a QSAR/QSPR data structure containing structurally diverse compounds. Figure 4 shows the result of a linear regression analysis on the parabolic data set, and as expected a single line cannot adequately represent a parabola. Introduction of an additional line can quickly alleviate the problem by cutting the parabola around its axis of symmetry into two somewhat linear halves. Figure 5 shows how two lines generated by the GAS method describe the parabola. Two major trends associated with this artificial

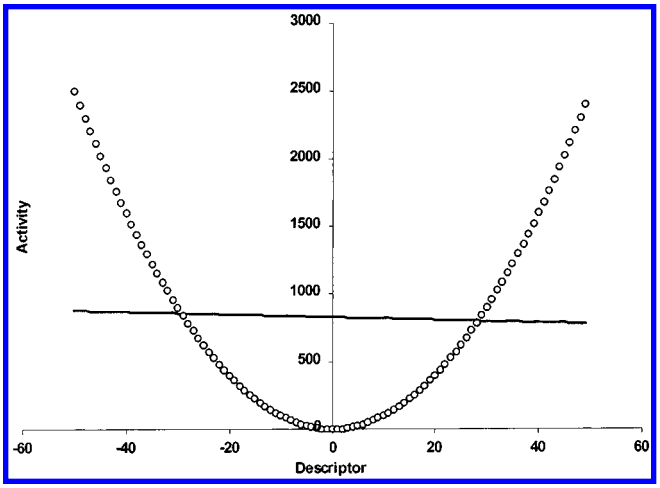

Figure 4. Description of the parabola data set using one equation model.

data set are explained by the two lines, one with a negative slope and the other with a positive slope. A better estimation of the parabola was obtained by isolating points near the vertex and describing them with a third line (Figure 6). Finally with four lines a very close approximation of the parabola can be obtained (Figure 7).

Subset Selection of the XLOGP Data Set. For comparison with the subset selection method, the XLOGP data set was first analyzed by MLR and the variable selection method. The result of the MLR analysis performed on the XLOGP data set is shown in Table 6. The r^2 values of 0.746 and 0.800 are obtained for the training and testing sets, respectively. The experimental and predicted log *P* values of the compounds in the test set are shown in Table 10 (Method F). Figures 8 and 9 show the correlation between the experimental and calculated log *P* values of compounds in the training and testing sets, respectively. The variable selection analysis was then performed to improve the MLR model. Table 7 shows the top 10 best models obtained. Among them, Parent 3 gave the best results with the r^2 value of 0.731 for compounds in the training set and the r^2 value of 0.766 for compounds in the test set. As shown by the ranges of r^2 values for compounds in the training and test sets (Table 7), performing the variable selection did not improve the quality of the MLR model. The correlation

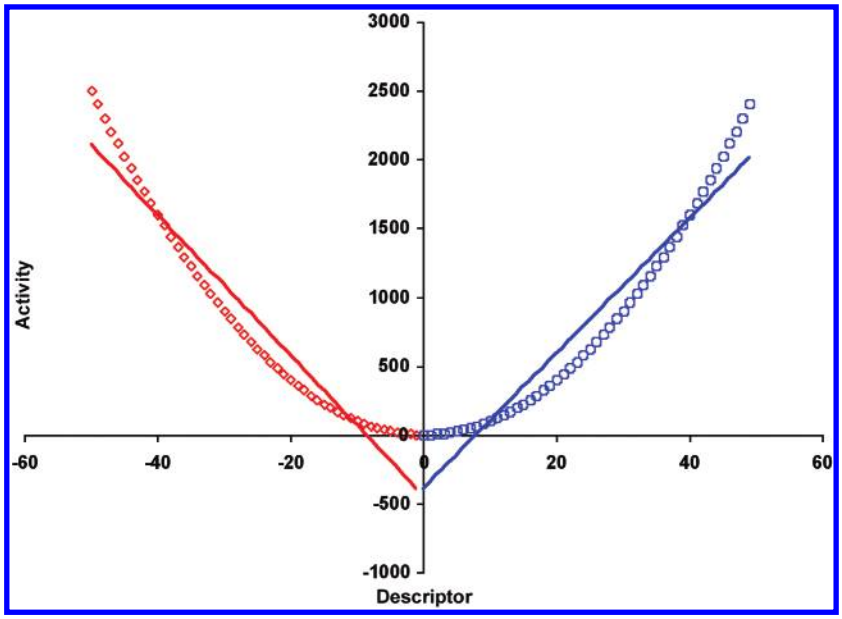

Figure 5. Description of the parabola data set using two equation models.

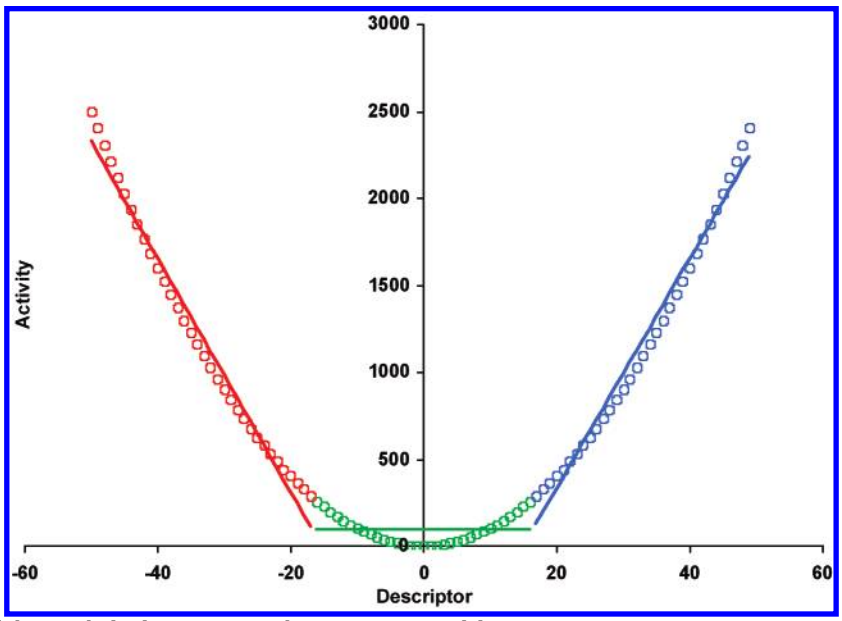

Figure 6. Description of the parabola data set using three equation models.

between the experimental and calculated log *P* values of compounds in the training set (calculated using the model obtained from Parent 3) is shown in Figure 10. The experimental and predicted log *P* values of compounds in the test set (calculated using the model obtained from Parent 3) are shown in Table 10 (Method G) and plotted in Figure 11. The result of the best parent obtained after the subset selection is shown in Table 8. Only two subset (or equation) models are available because the further increase in the number of subsets violates the "five compounds per variable" rule specified via the CVR parameter. The r^2 values of 0.853 and 0.839 are obtained for models 1 and 2, respectively. The predictive ability of each model was assessed by predicting log *P* values of compounds in Model 2 ($n = 908$) using Model 1 and predicting log *P* values of compounds in Model 1 ($n = 923$) using Model 2. Table 9 shows that the predictive r^2 values of 0.455 and 0.202 are obtained for predicting log *P* values of compounds in Models 1 and 2, respectively. The predictive r^2 values are substantially lower than r^2 values, suggesting that the subset selection is necessary to better correlate the data. Indeed, improved r^2 values (training set: 0.792; test set: 0.838) were obtained for compounds in the training and test sets when both models were used. Figure 12 shows the correlation of the actual and calculated log *P* values of compounds in the training set. The calculated values obtained using Models 1 and 2 were marked with "+" and "O", respectively. The correlation was tighter than ones generated with the MLR (Figure 8) and variable selection (Figure 10) methods. The actual and calculated log *P* values of compounds in the test set are shown in Table 10 (Method H) and plotted in Figure 13. The prediction of compounds in the test set using the subset selection method is shown to be clearly better than the prediction via MLR and variable selection methods (Methods F and G in Table 10). Table 10 also shows calculated log *P* values generated using other published prediction methods, including the XLOGP method (Method A), and the result of the subset selection method is shown to be comparable. In fact, if the

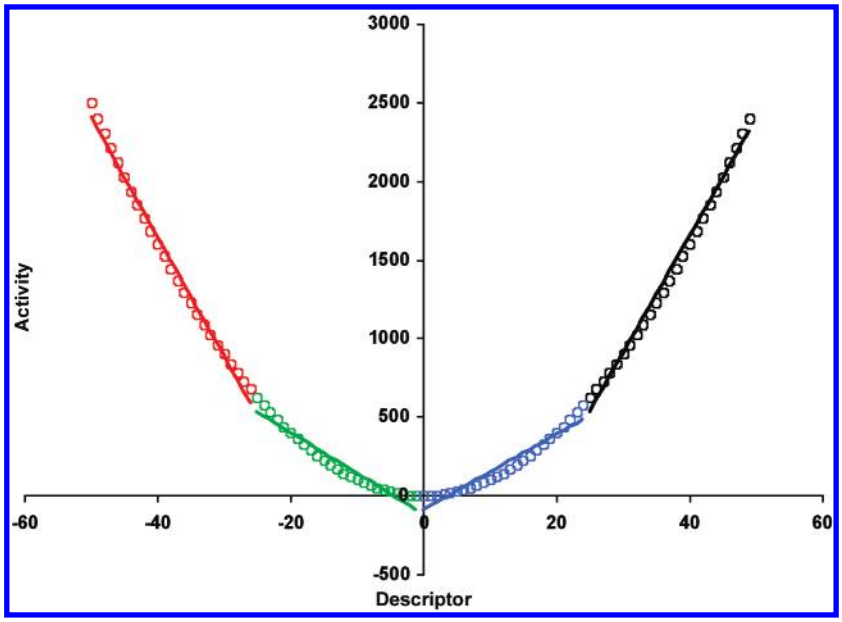

Figure 7. Description of the parabola data set using four equation models.

Table 6. Multiple Linear Regression Analysis of the XLOGP Data Set

	n	v	r^2	s	F	Q^2 [a]	SDEP[b]	r^2 (test set; N = 19)
MLR	1831	126	0.746	0.789	39.811	0.693	0.838	0.800

[a] Leave one out cross-validation. [b] Standard Error of Prediction.

Table 7. Variable Selection Analysis of the XLOGP Data Set (N = 1831)

parent	v	r^2	S	F	Q^2 [a]	SDEP[b]	r^2 (test set; N = 19)
1	36	0.730	0.793	134.902	0.711	0.812	0.760
2	35	0.729	0.794	138.216	0.710	0.814	0.764
3	37	0.731	0.792	131.642	0.711	0.812	0.766
4	36	0.730	0.793	134.800	0.711	0.813	0.756
5	37	0.731	0.792	131.599	0.711	0.812	0.757
6	37	0.731	0.792	131.577	0.711	0.812	0.753
7	37	0.731	0.792	131.566	0.711	0.812	0.757
8	36	0.730	0.793	134.705	0.710	0.813	0.766
9	37	0.731	0.792	131.533	0.711	0.812	0.757
10	36	0.730	0.793	134.674	0.710	0.813	0.762

[a] Leave one out cross-validation. [b] Standard Error of Prediction.

Table 8. Best Parent Obtained after the Subset Selection Analysis of the XLOGP Data Set (V = 126)

model	n	r^2	s	F	Q^2 [a]	SDEP[b]	r^2 (test set; N = 19)
1	923	0.853	0.647	36.534	0.811	0.680	
2	908	0.839	0.622	32.392	0.739	0.736	
1 & 2[c]	1831	0.792					0.838

[a] Leave one out cross-validation. [b] Standard Error of Prediction. [c] See prediction under the Methods section.

worst predicted compound, Cimetidine, is removed from the test set, the r^2 value of 0.934 is obtained for compounds in the test set using the subset selection method, compared to the r^2 value of 0.838. Finally, Table 11 shows the CPU time required to perform the various GAS runs. The subset selection method requires the most CPU time since the fitness of each subset has to be assessed for each parent.

Table 9. Predictive r^2 Values of Models 1 and 2 of the Best Parent Obtained after the Subset Selection Analysis of the XLOGP Data Set (V = 126)

	N	Predictive r^2
predict compounds in model 1 using model 2	923	0.455
predict compounds in model 2 using model 1	908	0.202

DISCUSSION

Traditionally, the optimization of a QSAR/QSPR model is performed at the descriptor level, and the compounds typically found in such a model tend to be structurally very similar. As the structural diversity in the training set increases, as in the case of compounds screened against HTS or ADME properties from a variety of therapeutic programs, the variable selection analysis alone may be inadequate to develop a good QSAR/QSPR model. In such cases, poor correlations do not necessarily mean that inappropriate descriptors were used or even that there is no correlation. It may mean that the data sets might contain two or more overlapping correlations complicating the construction of a single QSAR/QSPR model. Developing good QSAR/QSPR models probably depends less on the types of variables used (variable selection) and more on the types of compounds used (subset selection) in the data sets. In this paper, we report a genetic algorithm guided selection to address the subset selection problem. To illustrate the necessity of the subset selection, the MLR and variable selection analyses were also performed for a comparison. The validation of the variable selection method implemented in the GAS method was conducted using the Selwood data set[4] and compared with the published results (Table 5). The three best models using three variables reported by Rogers and Hopfinger[7] and Kubinyi[9] were identified using GAS (Table 4), and the variables found in the top 25 models encompass most of variables identified by other published variable selection methods (Table 5). The validation of the subset selection method implemented in GAS was performed using an artificial data set containing one descriptor (the visualization of the subset selection of a data set containing multidimensional points is difficult) and the XLOGP data set containing

Table 10. Comparison of Actual and Calculated logP Values Obtained by Different Calculation Methods

		calculation methods[a]							
drug	exp. logP[18]	A	B	C	D	E	F	G	H
atropine	1.83	2.29	2.21	1.88	1.32	0.03	1.15	1.27	1.83
chloramphenicol	1.14	1.46	1.23	0.32	0.69	−0.75	1.78	1.78	1.70
chlorothiazide	−0.24	−0.58	−0.36	−0.68	−1.24	−0.44	−0.06	0.00	0.17
chlorpromazine	5.19	4.91	3.77	5.10	5.20	3.89	5.48	5.19	5.19
cimetidine	0.40	0.20	0.82	0.63	0.21	3.33	−1.46	−1.38	−2.25
diazepam	2.99	2.98	3.36	3.18	3.32	1.23	3.50	3.77	3.16
diltiazem	2.70	3.14	2.67	4.53	3.55	1.96	4.20	4.40	2.37
diphenhydramine	3.27	3.74	3.26	3.41	2.93	3.35	3.65	3.76	3.51
flufenamic acid	5.25	4.45	3.86	5.81	5.58	5.16	3.90	3.69	4.29
haloperidol	4.30	4.35	4.01	3.57	3.52	3.43	3.18	3.11	4.24
imipramine	4.80	4.26	3.88	4.43	4.41	3.38	5.48	5.19	5.19
lidocaine	2.26	2.47	2.52	2.30	1.36	0.91	1.78	1.80	2.02
phenobarbital	1.47	1.77	0.78	1.23	1.37	1.29	1.46	1.38	1.88
phenytoin	2.47	2.23	1.80	2.76	2.09	2.01	2.32	2.26	2.33
procainamide	0.88	1.27	1.72	1.11	1.11	0.65	0.92	0.57	0.88
propranolol	2.98	2.98	2.53	3.46	2.75	2.15	2.37	2.38	2.55
tetracaine	3.73	2.73	2.64	3.55	3.65	2.90	3.83	3.51	3.09
trimethoprim	0.91	0.72	1.26	−0.07	0.66	0.57	1.06	0.59	1.28
verapamil	3.79	5.29	3.23	6.15	3.53	6.49	5.06	5.50	4.23
	r^2 =	0.887	0.870	0.841	0.941	0.543	0.800	0.766	0.838

[a] A − XLOGP;[17] B − Moriguchi;[19] C − Rekker;[19] D − Hansch-Leo;[19] E − Suzuki-Kudo;[19] F − multiple linear regression; G − variable selection (GAS; Parent 3); H − subset selection (GAS; Parent 1).

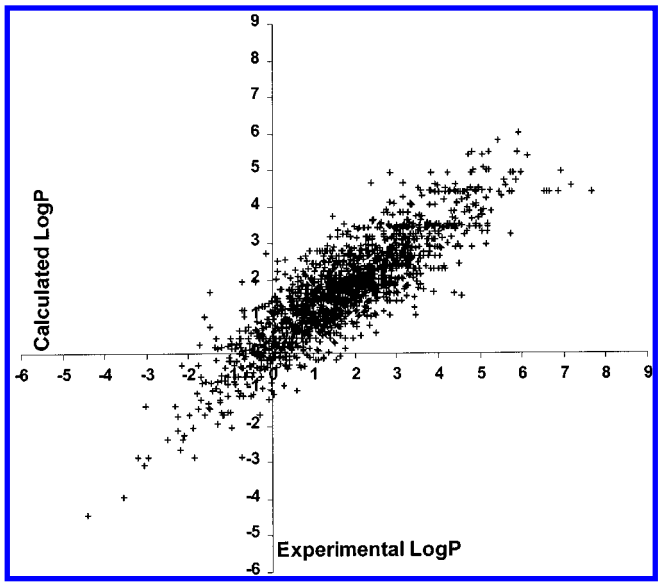

Figure 8. Actual and calculated activities generated using the multiple linear regression analysis of the XLOGP training set.

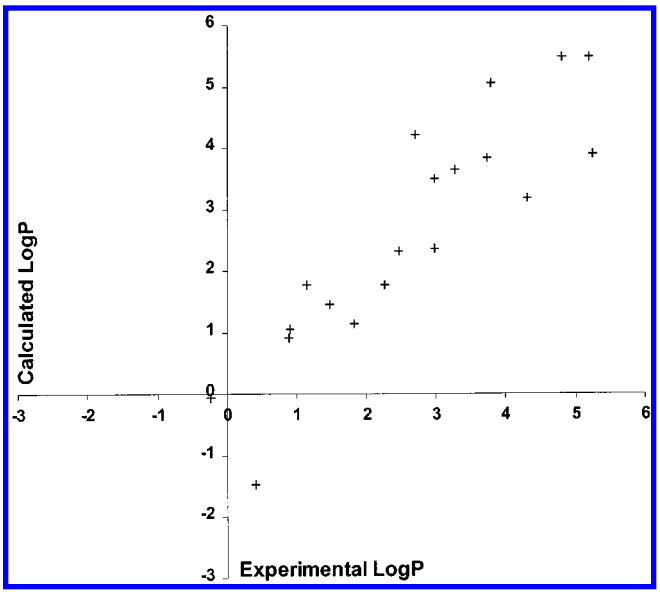

Figure 9. Actual and calculated activities generated using the multiple linear regression analysis of the XLOGP test set.

1831 compounds and 126 BIPF descriptors. The artificial data set was created simply using the equation of a parabola (eq 1). The test is to describe the nonlinearity of the parabolic data set (representing the complex nature of QSAR/QSPR model space) using the linear models generated via the subset selection method. As shown in Figure 4, correlating the nonlinear parabola using a line is difficult, but with subsequent addition of lines (subsets), the subset selection method appropriately represents the parabolic data points (Figures 4−7). The subset selection method was further examined using the XLOGP data set, which was analyzed via the MLR method and the variable selection method for comparison. The result of the MLR analysis was used as the bench mark for evaluating the variable and subset selections. The variable selection analysis of the XLOGP data set did not improve the MLR model. The best r^2 values of 0.731 (training set) and 0.766 (test set) are still lower than the r^2 values of 0.746 (training set) and 0.800 (test set) obtained using the MLR analysis (Tables 6 and 7). An improved model is obtained using the subset selection method; the r^2 values of 0.792 and 0.838 are obtained for the training and test sets, respectively (Table 8). The need for two subsets is demonstrated by the low r^2 for predicting log P of compounds in the first subset using the second model and compounds in the second subset using the first model (Table 9). If the subset model is able to predict the log *P* values of the compounds outside of the subset, this single model might be sufficient in describing the entire training set, and the subset selection would be unnecessary. Examining the quality of models generated via the subset selection method using the predictive r^2 value is, in fact, equivalent to determining the optimal number of subsets; the optimum number of subsets is determined by finding out how many subsets are required to describe the data without being redundant.

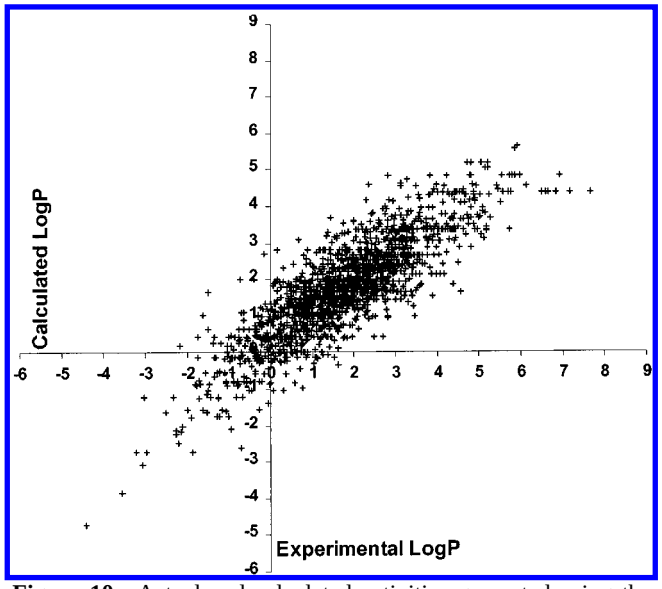

Figure 10. Actual and calculated activities generated using the variable selection (Parent 3) analysis of the XLOGP training set.

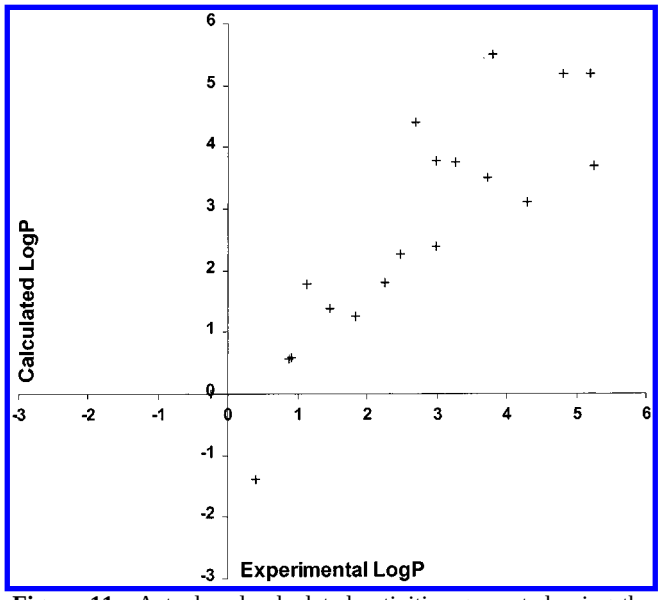

Figure 11. Actual and calculated activities generated using the variable selection (Parent 3) analysis of the XLOGP test set.

Table 11. CPU Time

method	CPU (s)
multiple linear regression[a]	9533
variable selection[b]	17964
subset selection[c]	91777

[a] Includes leave-one-out cross-validation. [b] Includes leave-one-out cross-validations of top 10 best parents in the final population. [c] Includes the leave-one-out cross-validation of the best parent in the final population.

In essence, the subset selection method can be considered as supervised clustering. It is supervised because the subset membership of compounds is based on the quality of a model that can be generated from the cluster (or subset in our case) in which the compounds belong to. Alternatively, a training set containing structurally diverse compounds can be clustered first, and each cluster can be used to build a model. The main problem of this approach is that the descriptors used to cluster the data set are weighed equally and may not be relevant to activity, whereas, in QSAR/QSPR analyses, the goal is to identify weights (or coefficients) associated with each descriptor. In the absence of any other information regarding the data set, clustering can be only performed using equally weighted descriptors. Since there is no way of determining the importance of each descriptor in advance, there is no guarantee that the subsets generated by cluster analysis would produce a good model.

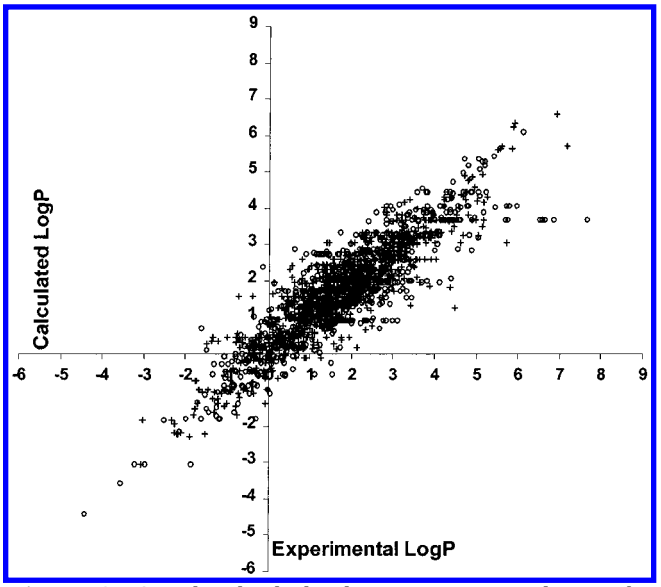

Figure 12. Actual and calculated activities generated using the subset selection (Parent 1) analysis of the XLOGP training set (+ − subset 1; O − subset 2).

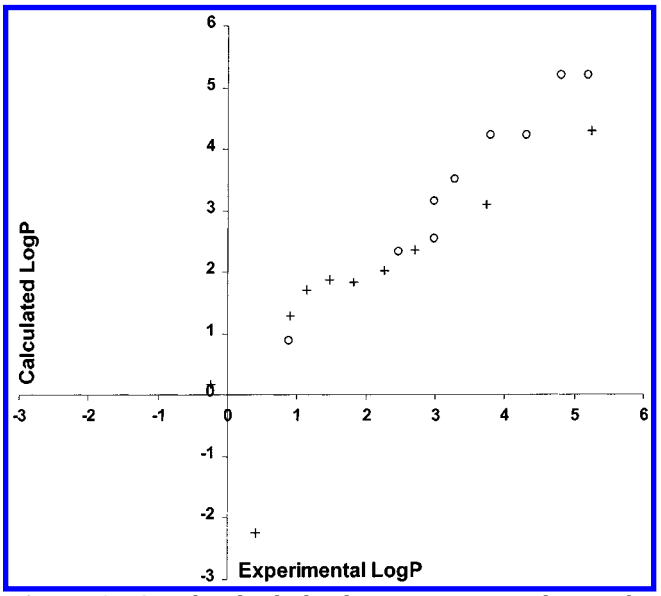

Figure 13. Actual and calculated activities generated using the subset selection (Parent 1) analysis of the XLOGP test set (+ − subset 1; O − subset 2).

CONCLUSIONS

The genetic algorithm guided selection method has been described. The method utilizes a simple encoding scheme which can represent both compounds and variables used to construct a QSAR/QSPR model. The variable selection method implemented in the GAS method has been tested and compared using the Selwood data set. The result of the variable selection showed that the method is comparable to

other published methods. The subset selection method implemented in the GAS method has been first tested using the parabolic data set to examine its ability to subset data points, and it was applied to the XLOGP data set. The analysis of the XLOGP data set showed that the subset selection method can be useful to improve a QSAR/QSPR model where the variable selection method fails. The method has been successfully used to develop an aqueous solubility prediction model[24] and is currently being applied to model other ADME properties. Performing subset selection has been found to be useful in developing models from a training set containing structurally diverse set of compounds. The method is susceptible to overfitting as the number of compounds decreases while the number of descriptors remains constant, but the initial application of variable selection and the CVR term in the fitting function can limit the risk of overfitting. Simultaneous optimization of both subsets and variables is currently being considered.

REFERENCES AND NOTES

(1) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, E.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(2) Hansch, C.; Fujita, T. $\rho-\sigma-\pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(3) Hansch, C. On the Structure of Medicinal Chemistry. *J. Med. Chem.* **1976**, *19*, 1−6.

(4) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure−Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.

(5) Wikel, J.; Dow, E. The Use of Neural Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−651.

(6) McFarland, J. W.; Gans, D. J. On Identifying Likely Determinants of Biological Activity in High Dimensional QSAR Problems. *Quant. Struct.-Act. Relat.* **1994**, *13*, 11−17.

(7) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(8) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.

(9) Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.

(10) Waller, C. L.; Bradley, M. P. Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure−Activity Relationship Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345−355.

(11) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative Structure−Activity Relationship Modeling of Dopamine D1 Antagonists Using Comparative Molecular Field Analysis, Genetic Algorithm-Partial Least Squares, and K Nearest Neighbor Methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(12) Weifan, Z.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the K−Nearest Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(13) Hoskuldsson, A. Variable and Subset Selection in PLS Regression. *Chemom. Intell. Lab. Syst.* **2001**, *55*, 23−38.

(14) Goldberg, D. E. *Genetic Algorithm in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(15) Holland, J. H. Genetic Algorithms. *Sci. Am.* **1992**, *267*, 66−72.

(16) Forrest, S. Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science* **1993**, *261*, 872−878.

(17) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.

(18) Hanch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: hydrophobic, electronic, and steric constants*; American Chemical Society: Washington, DC, 1995; Vol. 2.

(19) Moriguchi, I.; Hirono, S.; Nakagome, I.; Hirano, H. Comparison of Reliability of logP Values for Drugs Calculated by Several Methods. *Chem. Pharm. Bull.* **1994**, *42*, 976−978.

(20) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(21) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(22) Daylight Chemical Information Systems Inc., 27401 Los Altos, Suite #360, Mission Viejo, CA 92691.

(23) Waller, C. L.; Oprea, T. I.; Giolitti, A.; Marshall, G. R. Three-Dimensional QSAR of Human Immunodeficiency Virus(I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally-Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152−4160.

(24) Chen, X. Q.; Cho, S. J.; Li, Y.; Venkatesh, S. Prediction of Aqueous Solubility of Organic Compounds Using a Quantitative Structure−Property Relationship. *J. Pharm. Sci.* Submitted for publication.

CI010247V