

GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists¹

Kiyoshi Hasegawa,[‡] Yoshikatsu Miyashita,^{†,§} and Kimito Funatsu^{*,§}

Tokyo Research Laboratories, Kowa Co. Ltd., 17-43-2, Noguchi-cho Higashimurayama, Tokyo, 189, Japan,
and Department of Knowledge-Based Information Engineering, Toyohashi University of Technology,
Tempaku-cho, Toyohashi, 441, Japan

Received June 24, 1996[®]

The GAPLS (GA based PLS) program has been developed for variable selection in QSAR studies. The modified GA was employed to obtain a PLS model with high internal predictivity using a small number of variables. In order to show the performance of GAPLS for variable selection, the program was applied to the inhibitory activity of calcium channel antagonists. As a result, variables largely contributing to the inhibitory activity could be selected, and the structural requirements for the inhibitory activity could be estimated in an effective manner.

INTRODUCTION

QSARs are invaluable in the development of new agents as they permit interpretation of structure-activity data and prediction of compounds with some desirable pharmacological profiles.¹ QSARs are mathematical modeling of biological activity in terms of structural descriptors. The physico-chemical,² quantum-chemical,³ and topological parameters⁴ can be calculated by empirical and computational techniques and used as structural descriptors. Because the structural requirements for the biological activity are not known in advance, several types of structural descriptors should be considered in deriving a QSAR model.

In general, as the number of parameters increases, the quality of prediction of the QSAR model may decrease.⁵ Therefore, one has to attempt to select a set of parameters that produces the most predictive model. This problem is known as variable selection in QSAR studies.

Many groups have proposed different strategies for variable selection. J. H. Kalivas et al. have proposed the generalized simulated annealing (GSA) as a preferable method for variable selection.⁶ The GSA accepts or rejects a new combination of variables with probability, which is presented as a boltzmann function of change of predictability between the old and new QSAR models. H. Kubinyi has proposed mutation and selection uncover models (MUSEUM) to select important variables in antifilarial antimycin analogs.⁷ MUSEUM is the algorithm that a PLS model is evolved to the optimal one by random mutations and systematic addition or elimination of variables. S. Clementi et al. have proposed a variable selection procedure for PLS analysis, called the generating optimal linear PLS estimation (GOLPE).⁸ GOLPE preselects nonredundant variables using D-optimal design and then uses fractional factorial designs to create several PLS models with different combinations of variables. Variables significantly contributing to the

prediction are selected, while the others are eliminated. S. Wold et al. have proposed another variable selection procedure for PLS analysis, called the interactive variable selection (IVS).⁹ IVS is based on dimension-wise selection of each element in a PLS weight vector. The element in the PLS weight vector that is lower than the given limits in absolute value is set equal to zero, and a rescaling weight vector is used as the weight vector in the PLS algorithm.

GAs are a relatively new optimization technique and can be used as an alternative method for variable selection. D. Rogers et al. have developed the genetic function approximation (GFA) algorithm for variable selection. GFA can derive not only better combinations of variables but also a better basis function using the GA and spline technique.¹⁰ R. Leardi et al. have employed GA for variable selection in the MLR model and shown its advantage as compared with the classical stepwise selection with F statistics.¹¹ Successively, R. Leardi has proposed a novel approach combining GA with PLS for variable selection.^{12a} Unfortunately, he mainly concentrated on a validation method (full and partial validation), and utility of the approach was not demonstrated in the paper.

We have noticed the Leardi's approach and have developed the program (GAPLS: GA-based PLS) for variable selection in QSAR studies.^{12b} The modified GA proposed by Leardi et al. was employed to obtain a PLS model with high internal predictivity using a small number of variables. In order to show the performance of GAPLS for variable selection, the program was applied to the inhibitory activity of calcium channel antagonists. As the result, variables strongly contributing to the inhibitory activity could be selected, and the structural requirements for the inhibitory activity could be estimated in an effective manner.

VARIABLE SELECTION

GAs. GAs are a simulated method by which a principle of a natural evolution in biology (the struggle for life) is modeled.¹³ The struggle for life is that the species having a high fitness under some environmental conditions can prevail in the next generation, and the best species may be reproduced by crossover together with random mutations of

* Corresponding author.

† Deceased.

‡ Kowa Co. Ltd.

§ Toyohashi University of Technology.

¹ Abbreviations: QSAR, quantitative structure–activity relationship; GA, genetic algorithm; PLS, partial least squares; MLR, multiple linear regression.

[®] Abstract published in *Advance ACS Abstracts*, February 1, 1997.

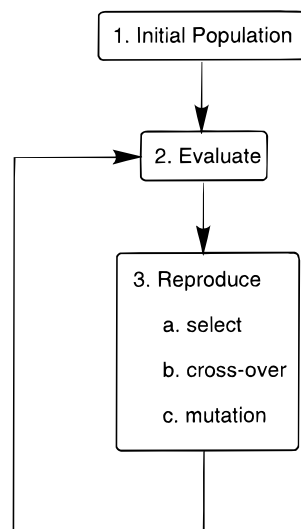


Figure 1. The flow chart of the GA scheme.

genes (chromosome) in the surviving ones. GAs were the first mathematically developed by J. H. Holland¹³ and later were applied as an optimization technique in several scientific fields.^{14,15} In GA for variable selection, the chromosome and its fitness in the species represent a set of variables and predictivity of the derived QSAR model, respectively.

Figure 1 shows a GA scheme for variable selection. The GA consists of three basic steps: (1) An initial population of chromosomes is created. The number of population (N_p) is dependent on dimensions of application problems. Each chromosome is represented by a binary bit string. Bit "1" denotes a selection of the corresponding variable, and bit "0" denotes a nonselection. The values of a binary bit are determined in a random way (probability of initial variable selection (p_i)). (2) A fitness of each chromosome in the population is evaluated by predictivity of the QSAR model derived from the binary bit string. The squared predictive correlation coefficient (r_{pred}^2) by the leave-one-out procedure was used as predictivity of the QSAR model. For the definition of r_{pred}^2 , see the section of PLS described below. (3) The population of chromosomes in the next generation is reproduced. This step can be divided into three operations: selection, crossover, and mutation.

First, a better chromosome is selected according to their fitness (the r_{pred}^2 value). The probability of a particular chromosome being selected (p_s) is computed from the relative value of fitness in the population, and it is compared to a random number (r). If the p_s value is higher than the r value, the chromosome is selected for the next crossover operation.

$$\text{selection probability } (p_s) = \text{fitness} / \text{sum of fitness} \quad (1)$$

This operation is repeated N_p times, and better chromosomes are selected.

Second, for any pairs of chromosomes, a random number is drawn to decide whether the crossover operation is undertaken or not. Crossover is an operation that a pair of chromosomes is individually divided, mutually exchanged, and merged. The probability of crossover (p_r) is set at the middle level so that most of pairs of chromosomes undergo this operation. If the crossover operation is to happen, a crossover point (c_{point}) ranging from 1 to $g - 1$ is determined from a random number (r), with g being the length of the

binary bit string, the number of variables.

$$c_{\text{point}} = \text{integer}(r(g - 1)) + 1 \quad (2)$$

Third, a binary bit string is mutated. For each binary bit, a random number is drawn to decide whether its bit has to be changed or not. The probability of mutation (p_m) is set at low level in the way that the overall fitness in the population of chromosomes can be improved successively.

The evaluation step (step 2) and reproduction step (step 3) in the GA scheme are continued until the number of the above repetitions is reached the designated number of generation (N_g).

Protection of More Informative Chromosomes. The importance of a chromosome is determined both by the prediction it gives and the number of variables it uses. That is, the chromosome with high interval predictivity using a small number of variables can be considered to be informative. A chromosome with k variables is defined as the most informative one when it gives the best prediction among all the chromosomes with at most k variables. Because the more informative chromosome has probability to produce an useful QSAR model (predictive and easily interpretable model), such a chromosome should not be eliminated from the population of chromosomes in the GA scheme.

In order to protect more informative chromosomes, an extra step, which decides whether a new chromosome replaces the old one or not, was introduced after the reproduction step (step 3) in the GA scheme according to the Leardi's algorithm.¹¹ The rule in the extra step was defined as follows: (a) If a new chromosome is a protected one, it is replaced with the least-fit nonprotected chromosome. (b) If a new chromosome is a nonprotected one, it is replaced with the old one only if its predictivity is higher than the predictivity of the least-fit nonprotected chromosome. Otherwise, the new chromosome is rejected.

By adding the extra step, the final population of chromosomes is composed of the most-fit chromosomes and the highly informative chromosomes. The structural requirements for the biological activity can be estimated from the selected variables that the highly informative chromosome indicates.

PLS. PLS was employed as a statistical method for the evaluation of fitness in the GA scheme. PLS has been widely employed to solve a multivariate calibration in analytical chemistry¹⁶ and the multivariate structure-activity relationships in QSAR.^{17,18}

The PLS model is derived in a principal component-like expression.¹⁹ The independent variables (X) and dependent variable (y) are modeled by a latent variable t . The X and y blocks are

$$X = \bar{X} + \sum_{h=1}^A t_h p_h^T + E \quad (3)$$

$$y = \bar{y} + \sum_{h=1}^A t_h q_h + f \quad (4)$$

where \bar{X} and \bar{y} are the corresponding means; p_h and q_h are the loading for the X and y blocks in the h th component, respectively; E and f are model residuals of X and y ,

respectively. A is the optimum number of components for the PLS model.

The latent variable t_h is expressed using a linear combination of X and w_h :

$$t_h = \left(X - \bar{X} - \sum_{i=1}^{h-1} t_i p_i^T \right) w_h \quad (5)$$

If eq 5 is substituted in eq 4, then a MLR-like model is obtained

$$y = \bar{y} + s(X - \bar{X}) \quad (6)$$

where s is a coefficient for the PLS model. The MLR-like model can provide insight in terms of original independent variables X .²⁰ In QSAR studies, X is a matrix consisting of structural descriptors and y is a vector consisting of biological activity.

In PLS analysis, a predictive PLS model can be obtained by selecting the optimum number of components (A) using a cross-validation technique.²¹ In the cross-validation technique, one or more samples in the data set are omitted, and the rederived PLS model is used to predict the biological activity of the omitted samples. This process is repeated until the biological activity of all samples in the data set has been predicted once. In this analysis, the leave-one-out procedure was employed as the cross-validation technique. The cross-validation technique gives a squared predictive correlation coefficient r_{pred}^2 as a statistical index. r_{pred}^2 is defined as follows

$$r_{\text{pred}}^2 = 1 - \frac{\sum_i (y_{i,\text{obs}} - y_{i,\text{pred}})^2}{\sum_i (y_{i,\text{obs}} - y_{\text{av}})^2} \quad (7)$$

where $y_{i,\text{obs}}$ and $y_{i,\text{pred}}$ are the observed and predicted activities for sample i , respectively. y_{av} is the averaged activity. The optimum number of components (A) is the PLS component which gives the highest r_{pred}^2 value in the cross-validation technique. The cross-validation technique is one of the powerful diagnostic tools for eliminating chance correlations and avoiding overfitting in the PLS model.⁵

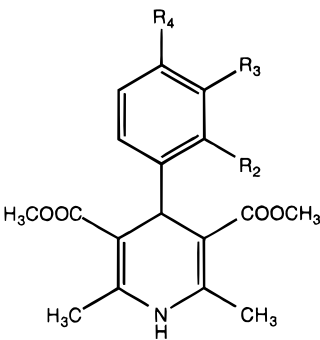
MATERIAL AND METHOD

Data Set. Thirty-five dihydropyridine (DHP) derivatives with the corresponding inhibitory activities were used as a data set for variable selection. The DHP derivatives are an important class of drugs known as calcium channel antagonists. The DHP derivatives act directly on the calcium channels and block the flux of calcium ions from the extracellular medium to the cell cytoplasm. Therefore, they can control the calcium-dependent biological events and treat cardiovascular diseases related to calcium channels.

The chemical structures and inhibitory activity of DHP derivatives are shown in Table 1. The activity data were taken from the study by Gaudio et al.²² Some DHP derivatives with two substituents at ortho or meta positions were eliminated from the data set because two bulky carboxyl esters on the dihydro pyridine ring may hinder a free rotation of the benzene ring, and two ortho or para substituents are not equivalent (Figure 1).

The $\log(1/\text{IC}_{50})$ values were used as dependent variables Y . IC_{50} is the molar concentration of a DHP derivative

Table 1. Chemical Structures and Observed and Calculated Inhibitory Activity of DHP Derivatives



no.	R_2	R_3	R_4	$\log(1/\text{IC}_{50})$	
				obs ^a	cal ^b
1 ^d	H	Br	H	8.89	7.72
2 ^c	CF ₃	H	H	8.82	8.65
3 ^d	Cl	H	H	8.66	8.37
4 ^c	H	NO ₂	H	8.40	8.04
5 ^d	CH=CH ₂	H	H	8.35	7.93
6 ^c	NO ₂	H	H	8.29	7.92
7 ^c	CH ₃	H	H	8.22	7.45
8 ^c	C ₂ H ₅	H	H	8.19	7.93
9 ^d	Br	H	H	8.12	8.56
10 ^c	CN	H	H	7.80	7.35
11 ^d	H	Cl	H	7.80	7.72
12 ^d	H	F	H	7.68	7.71
13 ^d	H	H	H	7.55	6.98
14 ^c	H	CN	H	7.46	7.00
15 ^c	H	I	H	7.38	7.57
16 ^d	F	H	H	7.37	7.72
17 ^c	I	H	H	7.33	8.76
18 ^d	OCH ₃	H	H	7.24	7.17
19 ^c	H	CF ₃	H	7.13	8.12
20 ^d	H	CH ₃	H	6.96	6.67
21 ^c	OC ₂ H ₅	H	H	6.96	7.55
22 ^d	H	OCH ₃	H	6.72	6.28
23 ^c	H	N(CH ₃) ₂	H	6.05	5.91
24 ^d	H	OH	H	6.00	6.50
25 ^c	H	NH ₂	H	5.70	5.23
26 ^c	H	OCOCH ₃	H	5.22	6.14
27 ^c	H	OCOPh	H	5.20	5.34
28 ^c	NH ₂	H	H	4.40	5.41
29 ^d	H	H	F	6.89	6.35
30 ^c	H	H	Br	5.40	5.10
31 ^c	H	H	I	4.64	4.67
32 ^c	H	H	NO ₂	5.50	5.51
33 ^c	H	H	N(CH ₃) ₂	4.00	5.42
34 ^c	H	H	CN	5.46	4.67
35 ^d	H	H	Cl	5.09	5.43

^a Observed inhibitory activity. ^b Calculated inhibitory activity by the PLS model using six variables. ^c Training set samples. ^d Test set samples.

necessary to inhibit 50% of the contraction of guinea pig ileum induced by methylfurmethide.

The structural descriptors describing each DHP derivative, which has to be selected by GA, is represented by a 12-dimensional vector x . Then, the independent variable X consists of 12 descriptors expressing the variation in 35 chemical structures of DHP derivatives.

$$x = (\pi(R_2), \sigma_m(R_2), B_1(R_2), L(R_2), \pi(R_3), \sigma_m(R_3), B_1(R_3), L(R_3), \pi(R_4), \sigma_m(R_4), B_1(R_4), L(R_4)) \quad (8)$$

The elements of the vector x are four physicochemical parameters ($\pi(R_2)$, $\sigma_m(R_2)$, $B_1(R_2)$, $L(R_2)$) for the substituent R_2 , four parameters ($\pi(R_3)$, $\sigma_m(R_3)$, $B_1(R_3)$, $L(R_3)$) for the

Table 2. Physicochemical Parameters Used for the PLS Analysis and GAPLS Analysis

no.	substituent	π	σ_m	B_1	L
1	Br	0.86	0.39	1.95	3.83
2	CF ₃	0.88	0.43	1.98	3.30
3	Cl	0.71	0.37	1.80	3.52
4	NO ₂	-0.28	0.71	1.70	3.44
5	CH=CH ₂	0.82	0.05	1.60	4.29
6	CH ₃	0.56	-0.07	1.52	3.00
7	C ₂ H ₅	1.02	-0.07	1.52	4.11
8	CN	-0.57	0.56	1.60	4.23
9	F	0.14	0.34	1.35	2.65
10	H	0.00	0.00	1.00	2.06
11	I	1.12	0.35	2.15	4.23
12	OCH ₃	-0.02	0.12	1.35	3.98
13	OC ₂ H ₅	0.38	0.10	1.35	4.92
14	N(CH ₃) ₂	0.18	-0.15	1.50	3.53
15	OH	-0.67	0.12	1.35	2.74
16	NH ₂	-1.23	-0.16	1.50	2.93
17	OCOCH ₃	-0.64	0.39	1.35	4.87
18	OCOPh	1.46	0.21	1.70	8.15

substituent R₃, and four parameters ($\pi(R_4)$, $\sigma_m(R_4)$, $B_1(R_4)$, $L(R_4)$) for the substituent R₄. π is a hydrophobic substituent constant, and σ_m is a Hammett σ constant for the meta position. B_1 and L are Verloop steric parameters, with B_1 being the smallest width and L being the length of the substituent. The values of physicochemical parameters have been quoted from the study of Gaudio et al.²² Table 2 shows the values of four physicochemical parameters of substituents.

GAPLS Computation. The GAPLS (GA-based PLS) program is written in FORTRAN language and is running on VAX workstation. The biological activity and physicochemical parameters were autoscaled to unit variance prior to the GAPLS computation. The PLS routine in the GAPLS program adopts standard PLS algorithm¹⁹ up to five numbers of components with the leave-one-out procedure.

The values of empirical parameters necessary for the GAPLS computation are as follows: The number of population (N_p) is 10, the probability of initial variable selection (p_i) is 0.5, the probability of crossover (p_r) is 0.5, the probability of mutation (p_m) is 0.1, and the number of generation (N_g) is 50. These values were determined to be optimal after several GAPLS computations with changing the values of empirical parameters.

RESULTS AND DISCUSSION

PLS Analysis. For a comparative study with the GAPLS analysis, the PLS analysis using 12 physicochemical parameters was carried out for the inhibitory activity of DHP derivatives. A four-component PLS model was derived by the leave-one-out procedure. The correlation coefficient, r , squared predictive correlation coefficient, r_{pred}^2 , and standard deviation, s , are 0.904, 0.571, and 0.623, respectively.

The PLS model was converted to the MLR-like model by the procedure discussed in the section of PLS.²⁰ Table 3 lists the values of the regression coefficient in the MLR-like model using 12 physicochemical parameters. Although some parameters have high regression coefficients, the structural requirements for the inhibitory activity are not so clear.

GAPLS Analysis. The GAPLS analysis was carried out to select the physicochemical parameters strongly contributing to the inhibitory activity of DHP derivatives. Table 4

Table 3. Regression Coefficients of the MLR-Like Model^a

no.	variable	model 1 ^b	model 2 ^c
1	$\pi(R_2)$	0.358	0.344
2	$\sigma_m(R_2)$	0.296	0.249
3	$B_1(R_2)$	-0.030	
4	$L(R_2)$	-0.008	
5	$\pi(R_3)$	0.150	0.248
6	$\sigma_m(R_3)$	0.407	0.418
7	$B_1(R_3)$	0.181	
8	$L(R_3)$	-0.567	-0.481
9	$\pi(R_4)$	-0.019	
10	$\sigma_m(R_4)$	0.181	
11	$B_1(R_4)$	-0.246	
12	$L(R_4)$	-0.377	-0.527

^a Coefficients represented by the autoscaled variables. ^b The MLR-like model using 12 variables. ^c The MLR-like model using six variables.

Table 4. Final Population of Chromosomes in the GAPLS Computation

chrom ^a	selected variables ^b	K ^c	A ^d	r_{pred}^2	protect ^e
1	1,2,5,6,8,12	6	4	0.685	1
2	1,2,5,6,8,9,12	7	4	0.684	0
3	1,2,5,6,8,11,12	7	4	0.682	0
4	1,2,5,6,8,9,11,12	8	4	0.681	0
5	1,2,5,6,7,8,9,12	8	4	0.679	0
6	1,2,4,5,6,8,12	7	5	0.678	0
7	1,2,5,6,7,8,12	7	4	0.677	0
8	1,2,5,6,7,8,11,12	8	4	0.674	0
9	1,2,4,5,6,8,9,12	8	4	0.673	0
10	1,2,6,7,8,12	6	3	0.673	0

^a The chromosome in the GA scheme. ^b The selected variables in the chromosome. The number corresponds with that of Table 3. ^c The number of variables. ^d The optimum number of PLS components. ^e "1" denotes the informative chromosome.

shows the final population of chromosomes in the GAPLS computation. In Table 4, only chromosome 1 (PLS model with four components using six variables) is an informative chromosome. Then, chromosome 1 was chosen for estimating the structural requirements for the inhibitory activity of DHP derivatives.

Chromosome 1 gave the highest predictive PLS model. The prediction of the PLS model was much improved by the GAPLS analysis from 0.571 to 0.685. The correlation coefficient r and standard deviation s are 0.893 and 0.655, respectively. The calculated inhibitory activity by the PLS model using six variables are shown in Table 1. Table 3 lists the values of the regression coefficient in the MLR-like model using six physicochemical parameters.

External Validations. Some researchers^{12a,23} have pointed out that special attention should be paid for choosing the measure of fitness for variable selection in GA. To check validity of the r_{pred}^2 value as the measure of fitness, two external validations using 12 and six physicochemical parameters were performed. We employed D-optimal criterion²⁴ to divide the data set into training and test sets. Table 1 shows the training and test sets determined by the D-optimal criterion.

The r value for the training set using 12 variables was 0.902 and r_{pred}^2 for the test set was 0.648. Similarly, the r value for the training set using six variables was 0.896, and r_{pred}^2 for the test set was 0.693. The predictivity of the PLS model is actually improved by variable selection, and the r_{pred}^2 value is considered to be a good measure of fitness in this GAPLS computation.

Structural Requirements. Although there was no experimental evidence that confirm the PLS model, it is useful to estimate the structural requirements for DHP derivatives by the model for drug design.

The selected physicochemical parameters at R₂ (ortho position) are π and σ_m . Hence, an ortho position is influenced by the hydrophobic and electronic factors. It seems that hydrophobic and electron-withdrawing substituents have advantage for the inhibitory activity at ortho position. The selected physicochemical parameters at R₃ (meta position) are π , σ_m , and L. A meta region seems to be a mixed region for the hydrophobic/electronic/steric parameters as Gaudio et al.²² emphasized. The interaction of the meta substituent with the receptor is rather complex, and further structure-activity data may be required to clarify its interaction mode. The selected physicochemical parameters at R₄ (para position) is L. A para position is specially affected by the steric parameter. It seems that a small pocket in the receptor exists and that the para substituent with a limited size of volume is required for the inhibitory activity.

The present structural requirements of DHP derivatives nicely corresponds to that derived by Gaudio et al.²² This suggests that the GAPLS program is useful to estimate the structural requirements in an effective manner.

CONCLUSION

In the present study, we have developed the GAPLS program for variable selection in QSAR studies. The modified GA proposed by Leardi was employed so that the more informative PLS models (chromosomes) were selected with high probability. By applying the program to the inhibitory activity of calcium channel antagonists, the structural requirements for the inhibitory activity could be estimated in an effective manner. The structural requirements nicely corresponds to that derived by the MLR analysis, and the utility of GAPLS was demonstrated.

ACKNOWLEDGMENT

The thorough and helpful comments of the reviewers on a preliminary version of the article are gratefully acknowledged. The comments of Dr. Carlos A del Carpio at Toyohashi University of Technology also helped to clarify the article.

REFERENCES AND NOTES

- (1) Martin, Y. C. *Quantitative Drug Design*; Marcel Dekker: New York, 1978.
- (2) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons: New York, 1979.
- (3) Nilson, L. M.; Carter, P. E.; Sterner, O.; Liljefors, T. Structure-Activity Relationships for Unsaturated Dialdehydes. 2. A PLS Correlation of Theoretical Descriptors for Six Compounds with Mutagenic Activity in the Ames Salmonella Assay. *Quant. Struct.-Act. Relat.* **1988**, *7*, 84–91.
- (4) Balaban, A. T.; Niculescu-Duvas, I.; Simon, Z. Topological aspects in QSAR for biologically active molecules. *Acta Pharm. Jugosl.* **1987**, *37*, 7–36.
- (5) Miyashita, Y.; Li, Z.; Sasaki, S. Chemical Pattern Recognition and Multivariate Analysis for QSAR studies. *Trends Anal. Chem.* **1993**, *12*, 50–60.
- (6) Sutter, J. M.; Kalivas, J. H. Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. *Microchem. J.* **1993**, *47*, 60–66.
- (7) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285–294.
- (8) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generalizing Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
- (9) Lindgren, F.; Geladi, P.; Rannar, S.; Wold, S. Interactive Variable Selection (IVS) for PLS. Part 1: Theory and Algorithms. *J. Chemom.* **1994**, *8*, 349–363.
- (10) 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.
- (11) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms for a strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267–281.
- (12) (a) Leardi, R. Application of genetic algorithms to Feature Selection Under full validation conditions and to outlier detection. *J. Chemom.* **1994**, *8*, 65–79. (b) During a subscription of this manuscript, another study for variable selection using GA and PLS has been made: Roger, D.; Dunn III, W. J. Genetic Partial Least Squares. *J. Comput.-Aided Mol. Des.* In Press.
- (13) Goldberg, D. E. *Genetic Algorithms in Search, Optimization & Machine Learning*; Addison-Wesley: New York, 1989.
- (14) Lucasius, C. B.; Kateman, G. Understanding and using genetic algorithms. Part 1. Concepts, properties and context. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 1–33.
- (15) Lucasius, C. B.; Kateman, G. Understanding and using genetic algorithms. Part 2. Representation, configuration and hybridization. *Chemom. Intell. Lab. Sys.* **1994**, *25*, 99–145.
- (16) Martens, H.; Noes, T. *Multivariate Calibration*; John Wiley & Sons: Chichester, 1989.
- (17) Hasegawa, K.; Deushi, T.; Yoshida, H.; Miyashita, Y.; Sasaki, S. Chemometric QSAR Studies of antifungal azoxy compounds. *J. Comput.-Aid. Mol. Des.* **1994**, *8*, 449–456.
- (18) Hasegawa, K.; Shigyo, H.; Sonoki, H.; Miyashita, Y.; Sasaki, S. Free-Wilson Discriminant Analysis of Antiarrhythmic Phenylpyridines Using PLS. *Quant. Struct.-Act. Relat.* **1995**, *14*, 344–347.
- (19) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
- (20) Kubinyi, H. *3D QSAR in Drug Design Theory, Methods and Applications*; ESCOM: Leiden, 1993.
- (21) van de Waterbeemd, H. *Chemometric Methods in Molecular Design, Method and Principles in Medicinal Chemistry*; Verlag Chemie: Weinheim, 1995; Vol. 2.
- (22) Gaudio, A. C.; Korolkovas, A.; Takahata, Y. Quantitative Structure-Activity Relationships for 1,4-Dihydropyridine Calcium Channel Antagonists (Nifedipine Analogues): A Quantum Chemical/Classical Approach. *J. Pharm. Sci.* **1994**, *83*, 1110–1115.
- (23) Rogers, D. Unpublished results.
- (24) Hasegawa, K.; Yokoo, N.; Watanabe, K.; Hirata, M.; Miyashita, Y.; Sasaki, S. Multivariate Free-Wilson analysis of α -chymotrypsin inhibitors using PLS. *Chemom. Intell. Lab. Sys.* **1996**, *33*, 63–69.

CI960047X