GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling

Toshiro Kimura,† Kiyoshi Hasegawa,‡ and Kimito Funatsu*,†

Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi 441, Japan, and Tokyo Research Laboratories, Kowa Company Ltd., Higashimurayama, Tokyo 189, Japan

Received June 17, 1997⊗

A novel approach using a genetic algorithm (GA) for variable selection in comparative molecular field analysis (CoMFA) was developed. This approach is named GA-based region selection (GARGS) since the regularly splitting regions in 3D space are used as variables instead of each field variable. GARGS was applied to the data set of polychlorinated dibenzofurans (PCDF) as a test example. The number of field variables was reduced from 1275 to 43, and the values of cross-validated r^2(q^2) indicating the internal predictivity of the model equation was increased from 0.88 to 0.95 by GARGS. The structural requirements for the PCDF molecules could be easily estimated from the coefficient contour maps of the simplified CoMFA model equation. These structural requirements were consistent with the result from the previous studies, and the utility of GARGS was demonstrated.

1. INTRODUCTION

Relationships between descriptors of chemical substances and their biological activities have been studied in the field of quantitative structure−activity relationships (QSAR). The main task of QSAR is to obtain a reliable model equation for prediction of the activities of new chemical substances.

In the classical QSAR studies, experimental physicochemical parameters of substituents in a chemical substance have been used as descriptors, and their effects on the biological activities have been investigated using multiple linear regression (MLR).[1] Although this approach has been applied to many QSAR problems,[2] it has limited utility for designing a new potent molecule because of no considerations of the 3D structure of molecules. In the late 1980's, a three-dimensional QSAR technique named comparative molecular field analysis (CoMFA) was introduced by Cramer et al.[3] In this method, steric and electrostatic interactions of probe atoms with the investigated molecules are used as descriptors, and the relationships between these 3D field descriptors and biological activities are modeled using partial least squares (PLS).[4,5] Since PLS is a robust regression method against collinearity among chemical descriptors, it can give the statistically significant model equations from the field variables and biological activities. The result of CoMFA can be displayed by computer graphics as the 3D coefficient contour maps, and the important regions for activity can be easily examined. Nowadays, CoMFA is widely used in the 3D-QSAR studies for drug design, and several successful results have been reported.[6]

In the QSAR studies, since the explicit mechanism for biological activity is not known in advance, many different variables should be investigated by trial and error. Generally, as the number of descriptors increases, the risk of chance correlation may increase.[7] Also, the model becomes complicated and its interpretation is difficult if we use many variables in modeling.[8] Therefore, it is important to select the meaningful variables to explain biological activities before or during modeling (variable selection). This is also true for CoMFA modeling.

Recently, several variable selection methods for CoMFA modeling have been proposed.[9,10] Lindgren et al. proposed the interactive variable selection (IVS) as a chemometric technique.[9] In their algorithm, variables are selected according to the weight value of the PLS model equation. There are two important steps in IVS: "inside-out" and "outside-in". In the inside-out step, every element of the PLS weight vector that is lower in absolute value than the given limit is set to zero. On the other hand, in the outside-in step, every element being larger than the given limit is set to zero. The rescaling vector from the above two steps is used as the weight vector in the PLS algorithm. Baroni et al. introduced the generating optimal linear PLS estimations (GOLPE) for the 3D-QSAR studies.[10] They performed the variable selection with two steps: (1) preliminary variable selection by means of D-optimal designs; (2) iterative evaluation of the effect of each variable on the model predictivity. In their procedure, the SDEP (standard deviation of the error of prediction) value is used to evaluate the model predictivity. The above two methods (IVS and GOLPE) appear to be useful for variable selection, but they have two main disadvantages because they belong to a systematic approach: (1) each choice heavily affects the following choices; (2) only the single variable combination is given to researchers.

A genetic algorithm (GA) is a novel optimization method based on the evolution process of beings. Because of its simplicity and effectiveness, GA has been applied to the

† Toyohashi University of Technology.
‡ Kowa Co. Ltd.
* To whom all correspondence should be addressed. E-mail address: funatsu@tutkie.tut.ac.jp.
⊗ Abstract published in *Advance ACS Abstracts,* December 1, 1997.

S0095-2338(97)00237-0 CCC: $15.00
Published on Web 03/23/1998

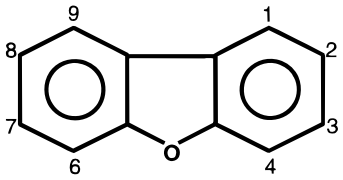

Figure 1. Chemical structure of the PCDF molecule.

various types of optimization problems in many scientific fields.[11−14] Several groups have combined GA with MLR and developed the novel variable selection methods in the QSAR studies.[15,16] Recently, our group has invented a promising variable selection method with GA and PLS (GA-based PLS: GAPLS).[17,18] In this method, PLS is employed as the statistical method and important variables are selected by GA using the cross-validated r^2 value of the PLS model. GAPLS was applied to some representative structure−activity data, and the predictive PLS models with important variables were obtained.

In the present study, we extended the GAPLS concept to deal with the 3D field variables in CoMFA. The 3D space in CoMFA was split into some domain regions, and they were used as variables instead of each field variable. This approach named GA-based region selection (GARGS) was applied to the data set of polychlorinated dibenzofurans (PCDF) which has been extensively studied by GOLPE.[10] Several splitting resolutions were examined to optimize the CoMFA model. By GARGS, the significantly improved CoMFA model equation was obtained as compared with the conventional CoMFA model in view of the cross-validated r^2 value and the number of field variables. As a result, the structural requirements for the PCDF molecules could be easily estimated from the coefficient contour maps of the simplified CoMFA model equation, which were consistent with the result from the previous studies.

2. MATERIALS AND METHODS

2.1. Data Set. Polychlorinated dibenzofurans (PCDF), analogous of polychlorinated dibenzo-*p*-dioxin (PCDD), are well-known environmental contaminants that have a strong toxicity for life. It has been known that the toxicity of the PCDF and PCDD molecules is related to the binding affinity for the aryl hydrocarbon (Ah) receptor. The relationship between the chemical structure of these molecules and Ah receptor binding affinity has been studied in the frame of QSAR methodologies.[19−21] Paso et al. reported the result of CoMFA modeling for Ah receptor binding of the PCDD and PCDF molecules.[21] In their paper, it was suggested that the Ah receptor binding is successfully described by the steric and electrostatic interactions.

In this study, we tried to construct the model for the PCDF data set in Baroni et al.[10] using the CoMFA method with the GARGS procedure. This data set is suitable for variable selection as a test example because it gives the messy coefficient contour maps by the conventional CoMFA modeling. Moreover, the result of GARGS can be compared with that of GOLPE, and the performance of GARGS will be checked with respect to the same data set. The chemical structure of the PCDF molecule is shown in Figure 1, and its Ah receptor binding affinity is listed in Table 1.

2.2. CoMFA. Comparative molecular field analysis (CoMFA) is one of the most powerful tools in the 3D-QSAR studies.[3] The CoMFA methodology is based on the assumption that the interactions between the molecule and enzyme is primarily noncovalent in nature. This method contains four fundamental steps: (1) generate the grid points around the investigated molecule with arbitrary spacing; (2) calculate the steric and electrostatic interaction between the investigated molecule and the probe atoms at each grid point; (3) construct the model equation between the interaction energies and biological activities using PLS; (4) display the coefficient contour maps of the PLS model equation in computer graphics.

Table 1. Data Set of Polychlorinated Dibenzofurans

ID	compound[a]	pEC_{50}	ID	compound[a]	pEC_{50}
1	1236-TCDF	−6.15	11	12467-PeCDF	−3.66
2	1248-TCDF	−5.22	12	12468-PeCDF	−5.15
3	2346-TCDF	−4.30	13	12478-PeCDF	−3.17
4	2347-TCDF	−2.39	14	12479-PeCDF	−2.72
5	2348-TCDF	−2.77	15	13478-PeCDF	−1.35
6	2368-TCDF	−4.15	16	23478-PeCDF	−0.55
7	2378-TCDF	−1.74	17	123478-HxCDF	−0.70
8	12348-PeCDF	−3.46	18	123678-HxCDF	−1.31
9	12378-PeCDF	−1.55	19	124678-HxCDF	−2.77
10	12379-PeCDF	−3.08	20	234678-HxCDF	−0.96

[a] Abbreviations: TCDF, tetrachlorodibenzofuran; PeCDF, pentachlorodibenzofurans; HxCDF, hexachlorodibenzofuran. For example, 1236-TCDF = 1,2,3,6-tetrachlorodibenzofuran.

2.3. PLS. Partial least squares (PLS) is a well-known regression method developed by H. Wold and S. Wold and belonging to the factor-based analysis similar to the concept such as principal component analysis (PCA) and principal component regression (PCR).[4,5] Nowadays, PLS has been widely used to solve the multivariate calibration in analytical chemistry and the multivariate structure−activity relationships in the QSAR studies.[8] In PLS, the relationship between parameters of the given samples (**X**) and the corresponding responses (**y**) is described using the linear model equation with latent variable **t**. The latent variable **t** is expressed by the linear combination of parameters, as shown in eq 1.

$$\mathbf{t} = \mathbf{X}\mathbf{w} \tag{1}$$

Here, **w** denotes the weight vector for calculation of the latent variable. Bilinear model equations are shown in eqs 2 and 3.

$$\mathbf{X} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p'}_i + \mathbf{E} \tag{2}$$

$$\mathbf{y} = \sum_{i=1}^{A} \mathbf{u}_i q_i + \mathbf{f} \tag{3}$$

In these model equations, **u** denotes the latent variable for the response variable. **p** and q are the loadings corresponding to **t** and **u**, respectively. **E** and **f** are the model residuals for **X** and **y**, respectively. A denotes the number of components in the PLS model equation.

The number of PLS components (A) is determined by means of a validation technique. Usually, the cross-validation technique named leave-one-out (LOO) is used. Three basic steps are contained in the LOO procedure: (1) extract one sample from the data set; (2) create the PLS model equation from the remaining samples; (3) predict the

responses for the extracted sample. These steps are repeatedly applied to obtain the predicted responses for all samples. After that, the predictive explained variance (denoted by q^2) indicating the internal prediction ability of the model equation is calculated according to eq 4.

$$q^2 = 1 - \sum_{\text{all samples}} (y_i - \hat{y}_i)^2 / \sum_{\text{all samples}} (y_i - \bar{y})^2 \quad (4)$$

Here, y_i denotes the observed response for sample i, $\bar{y}$ denotes the mean value of the response, and $\hat{y}_i$ denotes the predicted response for sample i. When q^2 is close to 1.0, it suggests that the model equation is most predictive. On the other hand, when the q^2 is small or negative, the model equation is useless for prediction. The number of PLS components (A) is set to the value which gives the maximum q^2 value.

2.4. GARGS. The genetic algorithm (GA) is one of the novel optimization methods based on the evolution process of beings.[11−16] In this method, parameters for the objective function are encoded into a binary string called a chromosome, and the chromosomes are evolved in the same manner as that of Darwin's theory.

In the present study, we invent a novel region selection method using the genetic algorithm named GA-based region selection (GARGS) for CoMFA modeling. In GARGS, the original CoMFA space is split into subboxes (regions) with arbitrary resolution along the *x*-, *y*-, and *z*-axes. Each region is encoded by a binary chromosome, and then the important regions are selected by the GA optimization procedure. The variable selection with regions instead of each field variable can be advantageous for several reasons:[22] (a) the structural change in the compounds is not reflected in a single field variable only but rather in a group of spatially contiguous field variables; (b) the CoMFA data matrices usually contain a large amount of variables, the result of PLS often becoming messy and very difficult to interpret; (c) the GA procedure takes much computing time, and its searching efficiency is low in the case with many field variables. The idea that regions instead of field variables are used as variables was taken from the literature.[23,24] A summary of the GARGS process is shown in Figure 2.

In the GA optimization procedure of GARGS, five fundamental steps are contained (see the GA optimization routine in Figure 2):

(1) Creation of the Initial Population. At the beginning, the initial population that contains the randomly generated chromosomes is created. The number of chromosomes in the population depends on the number of samples, the splitting resolution, and the complexity of the problem.

(2) Evaluation of the Fitness. Model equations with the selected variables according to the chromosome bit pattern are built using PLS, and their fitness value (predictivity of the model equation) is computed using the LOO procedure. The predictivity of the model equation is expressed in terms of the cross-validated r^2(q^2).

(3) Protection of the Informative Chromosome. Each chromosome with the least number of selected regions and the highest fitness is marked as an informative chromosome (protection). These chromosomes are kept from natural selection and crossover, to survive in the next generation preferentially. The idea of chromosome protection was taken from Leardi et al.[16]

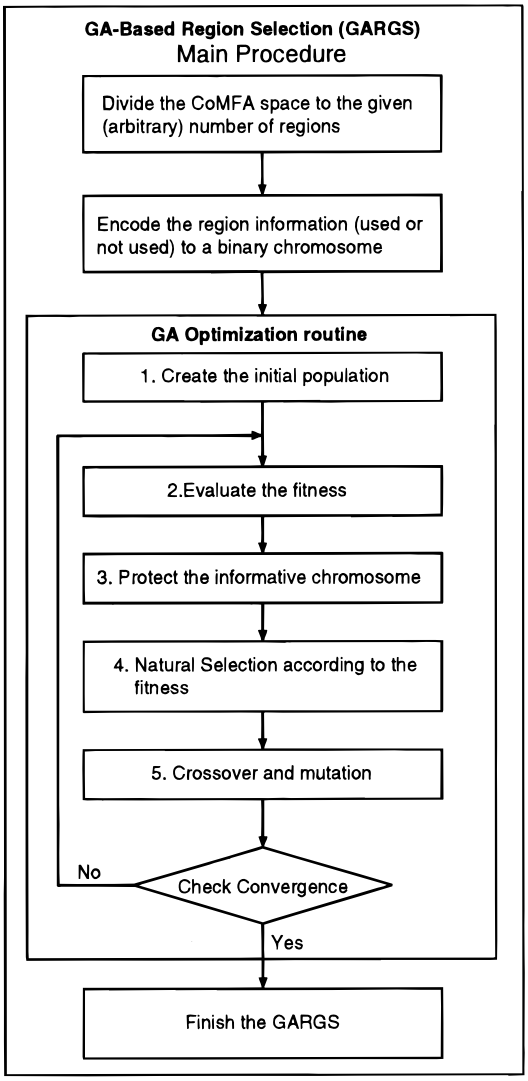

Figure 2. Summary of the GARGS procedure.

(4) Natural Selection According to the Fitness. The chromosomes with the higher fitness are selected from the population in arbitrary proportion. The proportion of selection from the population is set to 0.9.

(5) Crossover and Mutation. After natural selection, a new population consisting of chromosomes with the higher fitness is created. In the crossover procedure, new chromosomes are generated from a pair of randomly selected chromosomes (parent). Many methods have been proposed for the crossover technique.[13] The uniform crossover technique is employed and applied to 10 pairs of chromosomes in each iteration of generation. In the mutation procedure, the binary bit pattern in each chromosome is changed with a small probability. The probability of mutation is set to 0.01.

The four steps (from step 2 to 5) are repeatedly continued until the number of generation is reached to the given maximum. When the diversity of the population (the total amount of Hamming-distance between chromosomes) is smaller than the given criterion, the chromosomes are reconstructed by the random chromosomes.

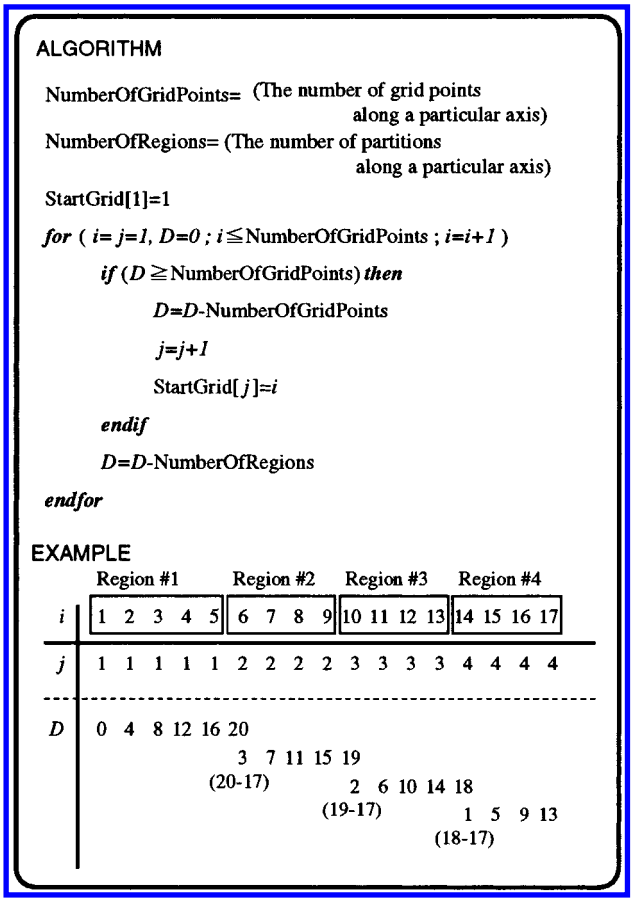

Figure 3. Region splitting algorithm using DDA and its example. This algorithm gives the boundary grid point in each region.

2.5. Region Splitting Strategy. As the beginning of GARGS, the CoMFA space should be split into some regions with arbitrary resolution. Digital differential analysis (DDA), one of the basic procedures for drawing lines on the screen in computer graphics,[25] was employed for splitting the CoMFA space. Figure 3 shows the DDA algorithm with its example for determining the boundary grid points in each region. In the algorithm shown in this figure, *i* and *j* represent the grid number and region number, respectively. *D* is a changing variable, and *D* is successively assigned to each grid point. If the value of *D* is greater than the number of grid points, the corresponding grid point assigned by the variable *i* is determined to be the boundary one. The example in Figure 3 shows the way by which the 17 grid points are split into 4 regions.

2.6. CoMFA Computation. All CoMFA computations were performed with the SYBYL software version 6.2[26] running on the Silicon Graphics INDY workstation. The electrostatic (Coulombic) and steric (Lennard-Jones) potentials were sampled for the grid points in 3D space around the set of molecules and evaluated as interactions with a probe sp^3 carbon atom having +1 charge. The distance-dependent dielectric model was used for the electrostatic potential calculation. The CoMFA grid spacing was 1 Å in all three dimensions within the defined space which extended beyond the van der Waals envelopes at all of the molecules. (*x*-axis, −8 to +8 Å; *y*-axis, −7 to +7 Å; *z*-axis, −4 to 0 Å). The dimension of the *z*-axis was set to the upper side only because of the symmetry of the PCDF molecules. From this setup, the total number of grid points is 1275 for steric and electrostatic fields, respectively. These definitions are the same as that of Baroni et al.[10] The parameters of GARGS are shown in Table 2.

Table 2. Parameters of GARGS

resolutions[a]	length of chromosome	no. of chromosomes	maximum generation
4-3-1	12	30	100
8-6-1	48	50	300
8-6-2	96	50	1000
8-6-5	240	50	2000

[a] *l*−*m*−*n* means that the number of partitions along the *x*-, *y*-, and *z*-axes are *l*, *m*, and *n*, respectively. See Figure 4.

Table 3. Results of the Conventional CoMFA Modeling

fields	no. of variables	no. of PLS components	r^2	q^2
steric and electrostatic	2550	4	0.957	0.837
electrostatic	1275	4	0.957	0.827
steric	1275	4	0.957	0.886

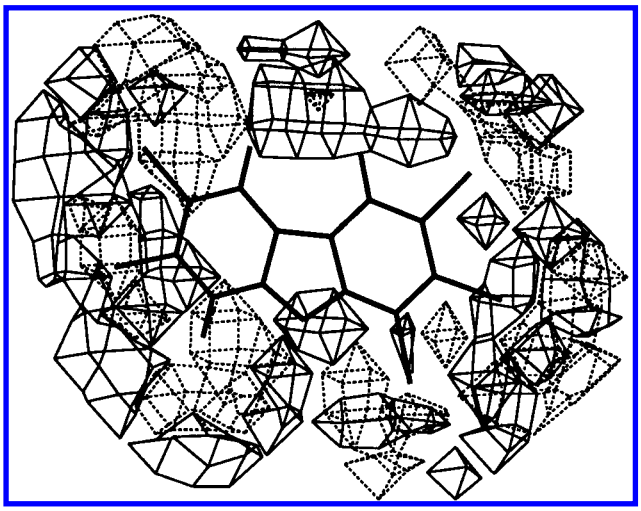

Figure 4. Coefficient contour maps of the conventional CoMFA model equation. The reference molecule is 23478-TCDF (1236-TCDF = 2,3,4,7,8-pentachlorodibenzofuran). The solid lines represent the coefficient values greater than 0.01, and dashed lines represent the coefficient values less than −0.01.

The GARGS program is written in C language. This program is implemented under the UNIX environment on an IBM-PC compatible microcomputer.

3. RESULTS AND DISCUSSION

3.1. Conventional CoMFA Modeling. For comparison with GARGS, conventional CoMFA modeling for the pEC_{50} values of PCDF was carried out with the steric and electrostatic fields. We constructed the CoMFA model equation with three combinations of fields: (1) steric plus electrostatic fields; (2) electrostatic field; (3) steric field. Results of the conventional CoMFA modeling was shown in Table 3. As shown in Table 3, the CoMFA model with steric field variables seems to be the best one. The conventional r^2 value is 0.96, and the cross-validated $r^2(q^2)$ value is 0.89 from this analysis. It can be considered that the pEC_{50} values of PCDF are largely explained by the steric field only. Figure 4 shows the coefficient contour maps of this model equation. This figure shows that the grid points having the large coefficient values are widely spread around

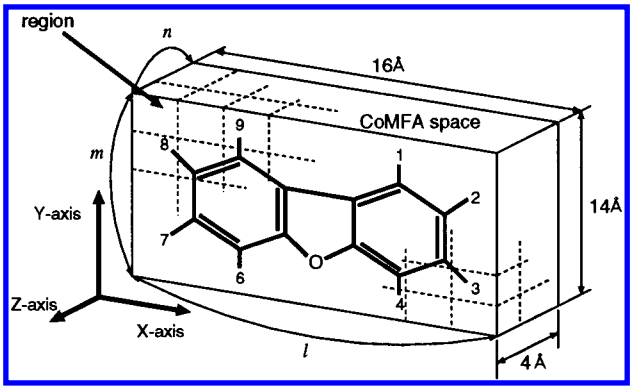

Figure 5. Region definition in the CoMFA space and CoMFA field computation settings of this study.

Table 4. Results of the CoMFA Modeling with GARGS

resolutions[a]		no. of selected regions (field variables)	no. of PLS component	r^2	q^2
4-3-1	I	5 (525)	5	0.950	0.906
	II	4 (425)	5	0.950	0.905
	III	3 (325)	5	0.953	0.903
8-6-1	I	8 (205)	5	0.970	0.946
	II	6 (175)	5	0.970	0.946
	III	4 (105)	5	0.967	0.940
8-6-2	I	10 (129)	5	0.972	0.952
	II	9 (117)	5	0.972	0.951
	III	8 (99)	5	0.972	0.951
8-6-5	I	10 (53)	5	0.973	0.952
	II	9 (49)	5	0.973	0.951
	III[b]	8 (43)	5	0.973	0.951

[a] See the footnote in Table 2. [b] Final CoMFA model.

the molecules. Since we cannot clearly find out the important regions of the PCDF molecules from the messy contour maps, variable selection by GARGS is necessary for this purpose.

3.2. CoMFA Modeling with GARGS. We defined the regions by the DDA algorithm and performed the CoMFA with the GARGS procedure. Several splitting resolutions (4-3-1, 8-6-1, 8-6-2, and 8-6-5) were examineed to optimize the CoMFA model. Here, *l*-*m*-*n* means that the number of partitions along the *x*-, *y*-, and *z*-axes are *l*, *m*, and *n*, respectively. (See Figure 5 as the schematic illustration). Only the steric field was used in this analysis because the pEC_{50} values for the PCDF molecules are largely explained by the steric field only, as discussed in the above section. Before region selection, regions having the small value of the standard deviations (0.5) were removed to avoid the noisy information derived from the theoretical computed field variables.

Results of CoMFA with GARGS are shown in Table 4. In this table, three high-ranked chromosomes with protection were marked as I, II, and III, respectively. Table 4 shows that the quality of the model (q^2) is improved compared with that of the conventional CoMFA model, and the higher resolution gives the better model equation. It is considered that the model equation derived from GARGS becomes more stable because only the informative grid points are selected by GARGS, and then the signal/noise ratio was increased.

The CoMFA model (III) with the resolution 8-6-5 was selected as the final one, because of the fewest variables and highest predictivity. The coefficient contour maps of the final CoMFA model is shown in Figure 6. The coefficient

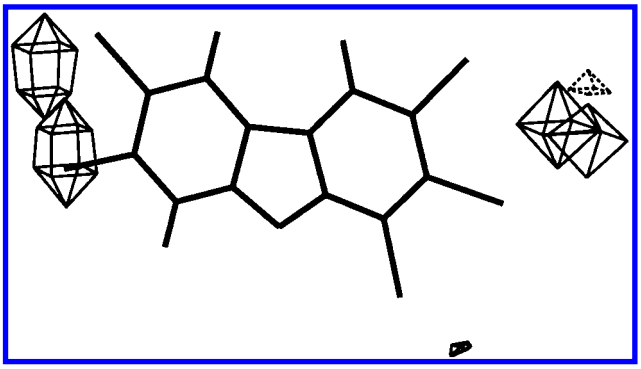

Figure 6. Coefficient contour maps of the final CoMFA model. The reference molecule is 23478-TCDF. The solid lines represent the coefficient values greater than 0.01, and dashed lines represent the coefficient values less than −0.01.

contour map was simplified with the use of GARGS, and the important regions for the PCDF molecules could be detected more clearly.

3.3. Structural Requirements. In Figure 6, it was shown that the important regions are located around the lateral positions of the PCDF molecules. In the previous structure−activity relationship study, it was proposed that the chlorine atoms at a lateral position are an advantage for the binding affinity of the PCDF molecules.[19−21] These structural requirements are consistent with the result of this CoMFA modeling. The coefficient contour map in Figure 6 clearly indicates the corresponding regions, and the utility of CoMFA with GARGS is validated.

In order to examine the CoMFA result in more detail, the interaction energies with the chlorine atom at each grid point were calculated from the final CoMFA model, and the 3D energy maps were displayed in computer graphics. Figure 7 shows the 3D energy maps with the threshold values of 0.5, 1.0, and 3.5 kcal/mol.

In the Figure 7A, the important regions around the position C-4 are shown. By comparison of A with B in Figure 7, it is easily estimated that the contribution of these regions on the binding affinity is approximately 0.5 in the pEC_{50} unit. The corresponding region for C-4 is not shown around the position C-6, although the PCDF molecule has the symmetrical shape. It may be caused from the fact that the C-4 substituted analogues are more active than the C-6 substituted analogues such as 1,2,3,6,7,8- and 1,2,3,4,7,8-hexachlorodibenzofuran (HxCDF).

Also, as shown in Figure 7C, the contributions of positions C-3 and C-7 to the binding affinity are larger than those of positions C-2 and C-8. This structural requirement is also consistent with the results in the previous studies.[19,20]

3.4. External Validation. Leardi et al. have pointed out that special attention should be paid for choosing the measure of fitness for variable selection in GA.[27] To check the validity of q^2 as the fitness, an external validation was performed. The data set was divided into two sets, named the training set and the test set, which are the same in Baroni et al.[10] The model equations of the training set with the final selected variables was used to predict the activities of the test set samples. Table 5 shows the results of external validation.

It is clear that the results from CoMFA with GARGS are better compared with those from the conventional CoMFA, based on the value of the prediction error sum of squares.

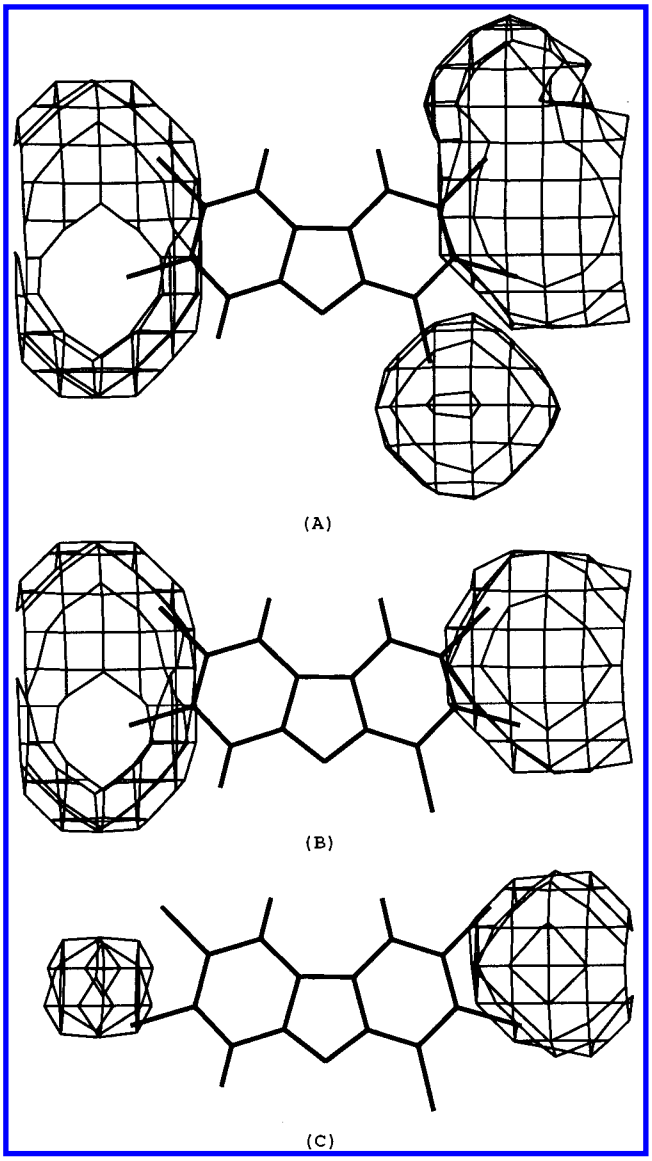

Figure 7. 3D energy maps with the chlorine atom, with the threshold values (A) 0.5, (B) 1.0, and (C) 3.5 kcal/mol, respectively.

Table 5. Results of External Validation

			Predicted pEC_{50}	
ID	Compound	Observed	Original model[a]	Final model[b]
2	1248-TCDF	−5.22	−4.39	−4.69
4	2347-TCDF	−2.39	−1.65	−2.07
11	12467-PeCDF	−3.66	−3.15	−3.90
16	23478-PeCDF	−0.55	−0.77	−1.01
18	123678-HxCDF	−1.31	−2.35	−1.89

[a] Conventional CoMFA model. [b] CoMFA model with GARGS.

(0.99 versus 2.63). These results are superior to those from the GOLPE method described in Baroni et al.[10] (0.99 versus 1.43). The predictivity of the CoMFA model is actually improved by variable selection, and the q^2 value is considered to be a good measure of fitness in GARGS.

4. CONCLUSION

In the present paper, we proposed a novel region selection method for CoMFA modeling named GARGS. We applied this approach to the data set of the PCDF molecules. CoMFA with GARGS gave more simple and significantly improved 3D-QSAR model equations compared with those from the conventional CoMFA. The structural requirements for the PCDF molecules could be easily estimated from the simplified 3D coefficient contour maps of the final CoMFA model. These structural requirements were consistent with the results from the previous studies, and the utility of GARGS was demonstrated.

CoMFA with GARGS is especially useful for modeling the data set in which molecules have many substitution positions within the relatively narrow 3D space. In this case, contributions of each substitution position for the activity will be mixed together, and the coefficient contour maps by the conventional CoMFA become messy and very difficult to interpret. GARGS can extract a subset of highly informative field variables as regions, and the chemically meaningful contour maps can be obtained. Moreover, it is also expected that GARGS can specify the molecular characteristics (steric, electrostatic, and hydrophobic character) at each region, and this matter will be further investigated in future work.

REFERENCES AND NOTES

(1) 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.
(2) Martin, Y. C. *Quantitative Drug Design*; Marcel Dekker, Inc.: New York, 1978.
(3) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(4) Geladi, P.; Kowalski, B. R. Partial Least Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(5) Höskluddson, A. PLS Regression Methods. *J. Chemom.* **1988**, *2*, 211−228.
(6) Kubinyi, H. *3D QSAR in Drug Design*; ESCOM Science Publishers: Leiden, The Netherlands, 1993.
(7) Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137−145.
(8) Miyashita, Y.; Li, Z.; Sasaki, S. Chemical Pattern Recognition and Multivariate Analysis for QSAR Studies. *Trends Anal. Chem.* **1993**, *12*, 50−60.
(9) Lindgren, F.; Geladi, P.; Rannar, S.; Wold, S. Interactive Variable Selection (IVS) for PLS. Part I: Theory and Algorithm. *J. Chemom.* **1994**, *8*, 349−363.
(10) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−12.
(11) Goldberg, D. E. *Genetic algorithm in search, optimization, and machine learning*; Addison-Wesley: New York, 1989.
(12) Davis, L. *Handbook of genetic algorithm*; Van Nostrand Reinhold: New York, 1991.
(13) Hibbert, D. B. Genetic Algorithm in Chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.
(14) Hibbert, D. B. Generation and display of chemical structures by genetic algorithms. *Chemom. Intell. Lab. Syst.* **1993**, *20*, 35−43.
(15) 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.
(16) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.
(17) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306−310.
(18) Hasegawa, K.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GAPLS and D-Optimal Designs for Predictive QSAR model. *J. Mol. Struct. (THEOCHEM)*, in press.
(19) Bandiera, S.; Sawyer, T.; Romkes, M.; Zmudzka, B.; Safe, L.; Mason, G.; Keys, B.; Safe, S. Polychlorinated Dibenzofurans (PCDFs): Effects of Structure on Binding to the 2,3,7,8-TCDD Cytosolic Receptor

Protein, AHH Induction and Toxicity. *Toxicology* **1984**, *32*, 131−144.
(20) Mason, G.; Sawyer, T.; Keys, B.; Bandiera, S.; Romkes, M.; Piskorska-Pliszczynska, J.; Zmudzka, B.; Safe, S. Polychlorinated Dibenzofurans (PCDFs): Correlation Between In Vivo and In Vitro Structure-Activity Relationships. *Toxicology* **1985**, *37*, 1−12.
(21) Paso, A.; Tuppurainen, K.; Ruuskanen, J.; Gynther, J. Binding of some dioxins and dibenzofurans to the Ah receptor. A QSAR model based on comparative molecular field analysis (CoMFA). *J. Mol. Struct. (THEOCHEM)* **1993**, *282*, 259−264.
(22) Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition: A New Way to Improve The Predictive Ability and Interpretability of Three-Dimensional Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1997**, *40*, 1455−1464.
(23) Norinder, U. Single and Domain Mode Variable Selection in 3D QSAR Applications. *J. Chemom.* **1996**, *10*, 95−105.
(24) Cho, S. J.; Tropsha, A. Cross-Validated R^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.
(25) Rogers, D. E. *Procedual elements for computer graphics*; McGraw-Hill Book Co.: New York, 1985.
(26) *SYBYL version 6.2 Theory Manual*; Tripos Inc.: St. Louis, MO, 1995.
(27) Leardi, R. Application of a Genetic Algorithm to Feature Selection under Full Validation Conditions and to Outlier Detection. *J. Chemom.* **1994**, *8*, 65−79.

CI970237N