GA Strategy for Variable Selection in QSAR Studies: Application of GA-Based Region Selection to a 3D-QSAR Study of Acetylcholinesterase Inhibitors

Kiyoshi Hasegawa[†]

Tokyo Research Laboratories, Kowa Co. Ltd., 2-17-43 Noguchi-cho, Higashimurayama, Tokyo, 189, Japan

Toshiro Kimura[‡] and Kimito Funatsu*

Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi, 441, Japan

Received May 5, 1998

Comparative molecular field analysis (CoMFA) with partial least squares (PLS) is one of the most frequently used tools in three-dimensional quantitative structure−activity relationships (3D-QSAR) studies. Although many successful CoMFA applications have proved the value of this approach, there are some problems in its proper application. Especially, the inability of PLS to handle the low signal-to-noise ratio (sample-to-variable ratio) has attracted much attention from QSAR researchers as an exciting research target, and several variable selection methods have been proposed. More recently, we have developed a novel variable selection method for CoMFA modeling (GARGS: genetic algorithm-based region selection), and its utility has been demonstrated in the previous paper (Kimura, T., et al. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 276−282). The purpose of this study is to evaluate whether GARGS can pinpoint known molecular interactions in 3D space. We have used a published set of acetylcholinesterase (AChE) inhibitors as a test example. By applying GARGS to a data set of AChE inhibitors, several improved models with high internal prediction and low number of field variables were obtained. External validation was performed to select a final model among them. The coefficient contour maps of the final GARGS model were compared with the properties of the active site in AChE and the consistency between them was evaluated.

1. INTRODUCTION

One important aim of drug design is to correlate the three-dimensional (3D) structures of molecules with their biological activities and to derive a predictive three-dimensional quantitative structure−activity relationship (3D-QSAR) model.[1] With a predictive 3D-QSAR model, medicinal chemists are able to design and predict biologically new potent molecules prior to synthesis. At present, one of the most frequently used tools for this task is Comparative Molecular Field Analysis (CoMFA).[2]

CoMFA attempts to establish a relationship between the biological activities and the steric/electrostatic properties of a set of compounds. After definition of a superposition rule for these compounds, the steric and electrostatic interaction energies with the probe atoms of each compound are calculated at each grid point in a box spanning 3D space. The outcome of this procedure is a matrix with many more columns (field interaction energies) than rows (compounds). To derive a linear equation from the resulting highly underdetermined matrix, a regression method called partial least squares (PLS) is applied.[3] PLS is not sensitive to collinearity of the underlying matrix as it operates with latent variables. PLS analysis is usually performed in combination with cross-validation in order to check for consistency of the model under consideration.[4] Cross-validation tests a model by omitting compounds, rederiving the model, and then predicting the activity of the omitted compounds, thus simulating the predictivity of the model. In this way the dimensionality of the model (the number of latent variables) is chosen according to its ability to predict the data rather than to fit the data. The results of CoMFA can be examined graphically as coefficient contour maps, and they can help in understanding the steric and electrostatic features of molecule correlated with high biological activity. In other words, these graphical representations can identify regions in 3D space that are favorable or unfavorable for ligand−receptor interactions.

There are several advantages when CoMFA is used in 3D-QSAR studies.[5] Since the steric and electrostatic field variables are calculated by the appropriate molecular mechanics parameters, no data point is excluded due to the lack of the tabulated substituent constants. Moreover, structurally heterogeneous compounds may be merged in the same model, because the field variables in CoMFA do not require a common molecular skeleton. As coupled with the available commercial software,[6] CoMFA has been applied to an ever-increasing number of QSAR data sets.

Despite the fact that many successful CoMFA applications have proved the value of this approach, especially in those cases where classical QSAR methods fail, there are some problems in its proper application.[7] The main problems that

* To whom all correspondence should be addressed. E-mail: funatsu@tutkie.tut.ac.jp.

[†] Current address: Nippon Roche Research Center, Nippon Roche K.K., 200 Kajiwara, Kamakura, Kanagawa, 247, Japan.

[‡] Current address: Sumisho Electronics Co. Ltd., 2-23 Shimomiyabi-cho, Shinjuku, Tokyo, 162, Japan.

10.1021/ci980088o CCC: $18.00
Published on Web 11/14/1998

have been encountered are the establishment of alignment/conformation rules and the inability of PLS to handle the low signal-to-noise ratio. The former point has been extensively discussed, and many promising methods have been proposed with the aid of X-ray crystallography of ligand−receptor complexes or empirical pharmacophore models derived from ligands.[8,9] The latter point, concerning the statistical limits of PLS, also attracts much attention from QSAR researchers because it has been found that the PLS method fails to unveil the variables correlated with activity when the signal-to-noise ratio (sample-to-variable ratio) is too low.[10] In most CoMFA studies, the structure of the receptor is unknown, and therefore, there is no well-defined way to assign the grid-spacing and grid-box dimension around molecules. As a consequence, a large number of ligand-probe interactions must be considered, most of that are irrelevant for explaining the biological activity. Therefore, selection of the most informative variables and elimination of background noise so as to increase the signal-to-noise ratio is a general problem, common to any 3D-QSAR method, including CoMFA with PLS.

In principle, there are two different strategies that can be pursued: one is a manipulation of the relative scaling of field variables and the other is a statistical selection of relevant field variables. The scaling of the field variables increases the weight of certain interactions that may dominate the regression model.[11] If these interactions are actually critical for activity, this can improve the signal-to-noise ratio. However, information about the important interactions is not known before analysis and in such cases it is likely that these matrix modifications will result in a merely mathematical model with no physical interpretation.

Selection of field variables seems to be a promising approach, and several variable selection methods for CoMFA modeling have been proposed. Lindgren et al.[12] proposed interactive variable selection (IVS) as a chemometric technique. In their algorithm, variables are selected according to the weight value in the PLS model. Baroni et al. proposed Generating Optimal Linear PLS Estimations (GOLPE) for 3D-QSAR studies.[13] GOLPE uses fractional factorial designs to create several PLS models with different combinations of variables and variables significantly contributing to the prediction are selected, while the others are eliminated. Cho et al. proposed cross-validated R^2 guided region selection (q^2-GRS) for CoMFA modeling.[14] q^2-GRS divides the CoMFA box into many small boxes, and a separate CoMFA is performed for each box. The boxes with associated q^2 values greater than a specified threshold value is selected for further analysis. More recently, an advanced variable selection method called genetic algorithm-based PLS (GA-PLS) has been developed in our group.[15,16] In GAPLS, PLS is employed as the statistical method, and variable combinations are selected by a genetic algorithm (GA) using the cross-validated r^2 value of the PLS model. With GAPLS, only a few significant variables are extracted from the large number of redundant variables. Furthermore, we have extended the GAPLS concept in order to deal with the 3D field variables in CoMFA.[17] An approach GA-based region selection (GARGS) uses domains of variables instead of each field variable. This strategy is advantageous for several reasons.[18,19] (1) Localized structural differences between compounds are not reflected in a single field variable but rather in a group of spatially contiguous field variables. (2) GA optimization requires much computing time, and its search efficiency becomes low in the case of large numbers of field variables. GARGS has been applied to a data set of polychlorinated dibenzofurans, and its utility has been demonstrated in a previous paper.[17]

The purpose of this study is to evaluate whether GARGS can pinpoint known molecular interactions in 3D space. We have used a published set of acetylcholinesterase (AChE) inhibitors[20] as a test example because the complex structure of ligand bound to enzyme has been solved by X-ray crystallography, and the complicated molecular conformation/alignment rule for CoMFA modeling need not be considered explicitly. By applying GARGS to the data set of AChE inhibitors, several improved models with high cross-validated r^2 value and low number of field variables were obtained. To select the final GARGS model, the data set was divided into training and test sets, and external validation was performed for each GARGS model. The selected model is significantly more predictive than the conventional CoMFA model with all field variables. The coefficient contour maps of the final GARGS model were compared with the properties of the active site in AChE, and the consistency between them was evaluated.

2. MATERIAL AND METHODS

2.1. Data Set. Sixty chemically diverse inhibitors for AChE, whose activity was measured by four different research groups, were used for this study. The data set has been compiled by Cho et al. and used for CoMFA modeling in their research course.[20] We selected this data set for two reasons. The series of molecules eliminates doubts arising from the alignment criteria and from the uncertainties about the conformations of the bound molecules because the data set includes three inhibitors, whose complex with AChE have been solved by X-ray crystallography. By using these three inhibitors as templates, it is not necessary to determine the complicated conformation/alignment rule for CoMFA modeling. Also, since the 3D structure of AChE is known, it is quite easy to compare the result of GARGS and the active site of AChE. The chemical structures of the AChE inhibitors are listed in Table 1 together with the inhibitory activities. The inhibitory activity is expressed as $-\log IC_{50}$; IC_{50} is the molar concentration of the inhibitor required to produce 50% inhibition.

2.2. Molecular Modeling. A detailed procedure for obtaining the 3D structures of compounds has been given in Cho's paper;[20] only a brief description will be given here.

The polypeptide backbones of three AChE/inhibitor complexes (THA: 9-amino-1,2,3,4-tetrahydroacridine, EDR: edrophonium, DME: decamethonium) were superimposed using an atom-based fit routine. The coordinates of the three inhibitors were extracted from their respective crystallographic complexes, and their geometry was optimized individually with the Tripos force field.[6] Each structure of the three inhibitors was used as a template onto which their analogues were superimposed. The initial analogue structures were constructed using the standard fragment library in SYBYL and optimized with the Tripos force field. After this initial process, all analogues were superimposed to the corresponding three templates and then field-fitted and

Table 1. Chemical Structures and Observed Inhibitory Activities of AChE Inhibitors

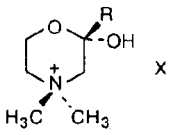
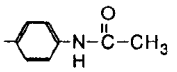
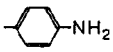
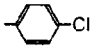
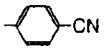
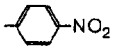
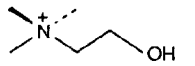
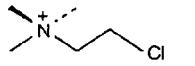
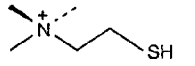
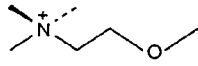
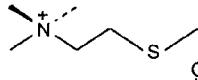
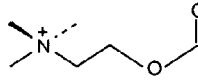
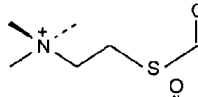
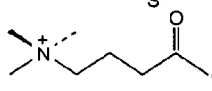
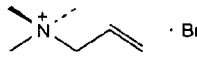
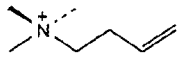
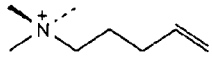
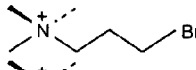
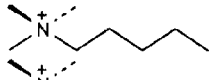
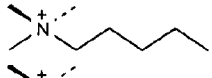
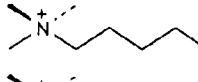
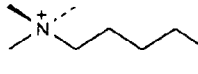
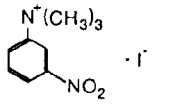
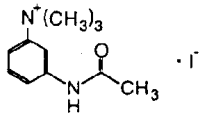
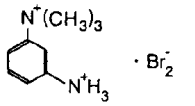
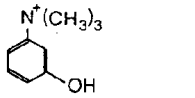
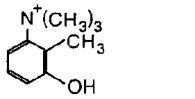
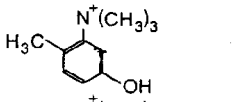
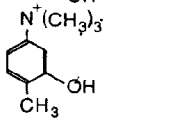
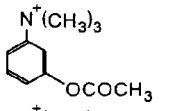
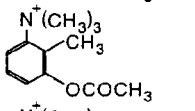
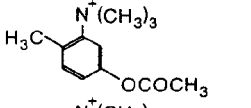
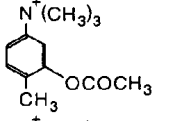
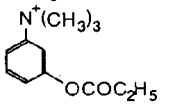
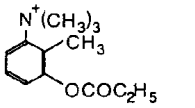
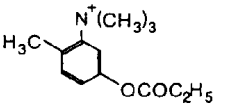
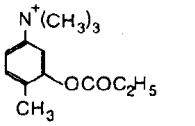
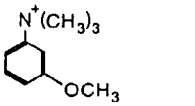
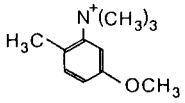
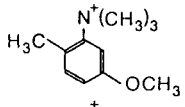
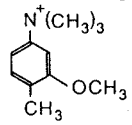
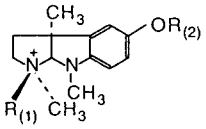
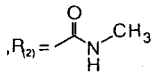
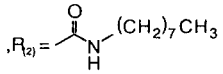
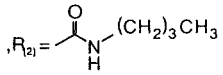
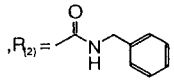
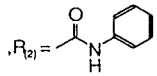
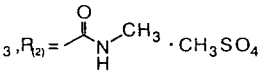
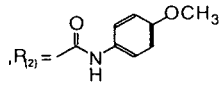
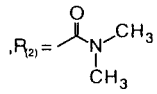
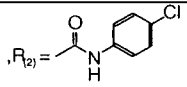
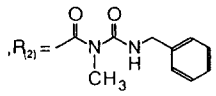
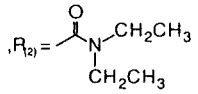
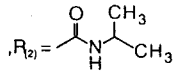
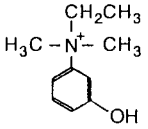
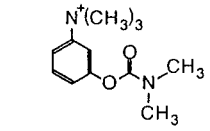
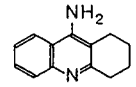
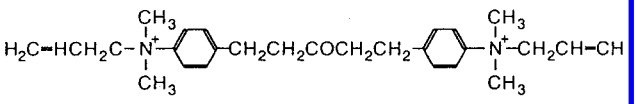
No.	Structure	$-\log IC_{50}$
		
1	R=  , X=Br^-	2.684
2	R=  , X=Br^-	3.161
3	R=  , X=Br^-	2.090
4	R=  , X=Br^-	1.936
5	R=  , X=Br^-	2.291
6	R=H , X=Cl^-	2.754
7	R=CH_3 , X=Br^-	2.762
8	R=$C(CH_3)$, X=Br^-	2.431
9	R=CF_3 , X=Br^-	2.417
10		2.521
11		2.622
12		3.357
13		2.936
14		2.821
15		3.640
16		3.900
17		4.072
18		2.535
19		3.072
20		2.947
21		4.056
22		3.224
23		3.327
24		3.272
25		3.088
26		3.202
27		3.000
28		3.717
29		6.012
30		3.851
31		5.548
32		5.507
33		5.521
34		4.389
35		5.272
36		6.181
37		5.224
38		3.123
39		4.161
40		5.424
41		3.224

Table 1. (Continued)

No.	Structure	$-\log IC_{50}$
42		3.622
43		3.462
44		3.914
		
45	$R_{(1)}$=H, $R_{(2)}$= 	7.244
46	$R_{(1)}$=H, $R_{(2)}$= 	6.959
47	$R_{(1)}$=H, $R_{(2)}$= 	6.818
48	$R_{(1)}$=H, $R_{(2)}$= 	6.337
49	$R_{(1)}$=H, $R_{(2)}$= 	6.456
50	$R_{(1)}$=CH_3, $R_{(2)}$= 	7.469
51	$R_{(1)}$=H, $R_{(2)}$= 	5.770
52	$R_{(1)}$=H, $R_{(2)}$= 	6.013
53	$R_{(1)}$=H, $R_{(2)}$= 	5.745
54	$R_{(1)}$=H, $R_{(2)}$= 	5.201
55	$R_{(1)}$=H, $R_{(2)}$= 	4.398
56	$R_{(1)}$=H, $R_{(2)}$= 	4.456
EDR		6.108
Neostigmine		7.041
THA		7.119
BW284C51		8.097

reoptimized with the field-fit option turned off. The atomic charges in each compound for the electrostatic interaction energy were calculated using the Gasteiger−Marsili method as implemented in SYBYL. The 3D coordinates of the aligned molecules with the atomic charges were kindly provided by Dr. Sung Jin Cho.

2.3. Conventional CoMFA. All CoMFA computations were performed on the aligned molecules, within the QSAR option of SYBYL version 6.2 running on a Silicon Graphics INDY workstation.[6] For all steps of CoMFA, the default SYBYL settings were used except otherwise noted. The CoMFA grid spacing was 2.0 Å in all three dimensions within the defined box, which was extended beyond the van der Waals envelopes of all molecules by at least 4.0 Å. From these definitions, the total number of grid points is 2856 (17*12*14) for both steric and electrostatic fields. The steric (Lennard-Jones) and electrostatic (Coulombic) interaction energies were calculated at a grid point in space with a C_{sp3} probe atom carrying a charge of 1.0. The energy cutoff value of 30 kcal/mol was applied for both the steric and electrostatic fields, and the minimum σ value was set to 2.0. These definitions were also used in the GARGS computation.

The field variable matrix was mean-centered and then block-scaled to give the same weight for the steric and electrostatic fields.[5] The resulting matrix was analyzed by PLS to build a multivariable regression model. PLS regression consists of three steps.[3] First, the data matrix is decomposed into a set of latent variables, taking into consideration the covariance between the field variables and biological activity. This results in a new orthogonal matrix of PLS scores. In the second step, the score matrix is regressed against the biological activity. Cross-validation tests are carried out to determine an optimum number of components.[4] In this study, the number of cross-validation groups was always equal to the number of compounds (leave-one-out technique), and the optimum number of components was chosen to be that giving the overall cross-validated r^2 maximum up to the 15 numbers of components. A final

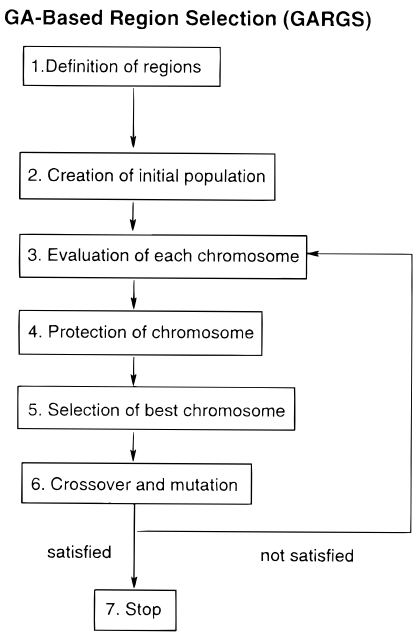


Figure 1. Flow chart of GARGS.

PLS analysis is performed using the chosen optimum number of components but with no cross-validation. This generates a fitted model of the entire data set with the r^2 value.

2.4. GARGS. GARGS is a sophisticated hybrid approach that combines GAs (a powerful optimization method) with PLS (a robust statistical method) for variable selection.[17] The related approaches have been already reported by several groups,[21−23] but GARGS is the first hybrid approach for variable selection of CoMFA modeling. GAs are optimization technique that mimic some of the processes observed in natural evolution.[24] A GA maintains a population of bit-strings (chromosomes) which represent candidate solutions that are in constant competition for survival. Using generalized evolutionary operations such as selection, crossover, and mutation, the population evolves dynamically toward improved solutions (higher fitness) in a number of cycles or generations; a summary of GARGS is shown in Figure 1. GARGS can be described by the following seven steps:

1. Definition of Regions. The original CoMFA box is divided into sub-boxes (regions) with arbitrary dimensions along the *x*-, *y*-, and *z*-axis. The resolution is expressed as a symbol ($l - m - n$) which means that the number of partitions along the *x*-, *y*-, and *z*-axis are *l*, *m*, and *n*, respectively. Digital differential analysis (DDA),[25] one of the basic techniques for drawing lines on the screen in computer graphics, is employed as the algorithm for this purpose.

2. Creation of Initial Population. The initial population of chromosomes is created by setting all bits in each chromosome to a random binary value (0 or 1); bit "1" denotes a selection of the corresponding region, and bit "0" denotes nonselection. The number of chromosomes in the population (N_p) is dependent on the number of samples, the number of regions, and complexity of the problem to be solved.[26]

3. Evaluation of Each Chromosome. The fitness of each chromosome is evaluated by the predictivity of the CoMFA model derived from the binary bit-string. The cross-validated r^2 value by the leave-one-out procedure is used as the index of prediction.

4. Protection of Chromosome. A chromosome with *k* variables is defined as an informative one when it gives the best fitness among all the chromosomes with at most *k* variables. The informative chromosome is protected and is kept during selection, crossover, and mutation to survive in the next generation preferentially. The detailed descriptions about chromosome protection are referred to the work by Leardi et al.[27]

5. Selection of Best Chromosomes. Chromosomes with highest fitness are selected from the population randomly (selection rate: P_s). The other chromosomes, which make up the next generation, are created by the following crossover and mutation step in order to ensure the diversity of population.

6. Crossover and Mutation. This step creates new chromosomes by the use of crossover and mutation operators. In crossover, a pair of randomly selected chromosomes is individually divided, mutually exchanged, and merged. The uniform crossover technique is employed and applied to predefined pairs of chromosomes in each generation (crossover frequency: F_c).[24] In mutation, the binary pattern in each chromosome is changed with a probability (mutation rate: P_m). The probability of mutation is set at low level such that the overall fitness of population can be improved successively. Steps 5 and 6 create all chromosomes used in the next generation.

7. Stop if a Halt Condition Is Satisfied Otherwise Go to Step 3. The cycle of evaluation, protection, selection, and crossover/mutation (steps 3−6) is repeated until a halt condition is satisfied. In this study, a maximum number of generations (N_g) is used as the halt condition, although other criteria such as the total Hamming-distance between chromosomes can also be used.[24]

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

3. RESULTS AND DISCUSSION

3.1. Conventional CoMFA Modeling. For a comparative study with GARGS, conventional CoMFA modeling with all steric and electrostatic fields was carried out. An initial PLS analysis was run to determine the optimum number of components using cross-validation. As the number of components was increased from one to five, the cross-validated r^2 values (denoted by q^2) were 0.482, 0.558, 0.651, 0.681, and 0.678, respectively. Thus, the optimum number of components was four from the cross-validation test. This number of components gave a model with r^2 and q^2 values of 0.892 and 0.681, respectively. A plot of the observed versus predicted inhibitory activities with four components is shown in Figure 2. Although the statistics of the conventional CoMFA model seem to be relatively satisfactory, some compounds such as nos. 26, 27, 55, 56, BW284C51, and Neostigmine were poorly predicted by this model. Furthermore the biased (unbalanced) coefficient contour map of steric field with only negative values was obtained at the 0.005 level which suggests the necessity of field variable selection.

3.2. GARGS Modeling. GARGS was applied to the data set of AChE inhibitors in order to reduce the number of field

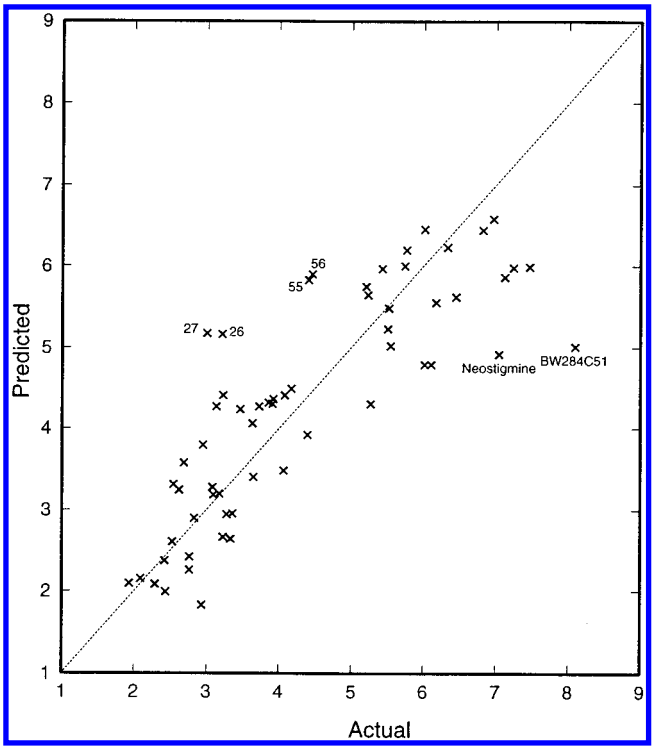


Figure 2. Plot of the observed and predicted inhibitory activities by the conventional CoMFA model.

Table 2. Parameters of GARGS

	N_p[a]	P_s[b]	F_c[c]	P_m[d]	N_g[e]
resolution 5-4-4	50	0.30	5	0.010	1000
resolution 8-6-7	100	0.30	15	0.001	3000

[a] Number of chromosomes. [b] Selection rate. [c] Crossover frequency. [d] Mutation rate. [e] Number of generations.

Table 3. Protected Chromosomes in GARGS with 5-4-4 Resolution

chromosome no.	no. of selected regions (field variables)	no. of components	r^2	q^2	PRESS
1	10(396)	4	0.864	0.800	16.72
2	9(348)	4	0.866	0.798	16.71
3	8(303)	4	0.850	0.794	15.98
4[a]	6(228)	4	0.852	0.790	15.44
5	5(192)	4	0.848	0.784	15.53
6	4(135)	5	0.871	0.771	22.76

[a] Represents the best protected chromosome.

variables and increase the prediction power of the model. Two splitting resolutions (5-4-4, 8-6-7), which approximately give the cubic regions, were examined in this study to avoid the biased results due to the shape of region. First, the case of GARGS with 5-4-4 resolution was investigated. The values of GA parameters affecting the performance of GARGS are summarized in Table 2; these values were empirically determined from experience in a previous study.[17] Table 3 shows the six high-ranked chromosomes with protection in the final GA population. All the protected chromosomes have higher q^2 values compared with those of conventional CoMFA model. There is a critical problem as to which model is selected from the GA population. The model with the highest q^2 value need not be best for prediction, because the q^2 value often leads to an overestimation of the model in the case of large data sets.[28] So, to select the best model, the data set was divided into training and test sets, and then external validation was performed for each protected model. The model with the best prediction for the test set was assumed to be the best GARGS model. Twelve (nos. 2, 4, 10, 17, 27, 36, 50, 55, EDR, Neostigmine, THA, and BW284C51) and the remaining 48 compounds were assigned as the test and training sets, respectively. By this assignment, the compounds in the test set have the maximum and minimum inhibitory activity in each chemical family. This allows the extrapolation test from the training set model. PRESS was used as the index of prediction for external validation. PRESS is the sum of the squared deviations between the predicted and measured biological activities. The result of external validation is shown in Table 3. Chromosome 4 has the minimum PRESS value, and this model was selected as the best one in the case of 5-4-4 resolution ($r^2 = 0.852$, $q^2 = 0.790$). The chromosome 4

Table 4. Protected Chromosomes in GARGS with 8-6-7 Resolution

chromosome no.	no. of selected regions (field variables)	no. of components	r^2	q^2	PRESS
1	20(172)	4	0.902	0.854	14.52
2	19(164)	4	0.903	0.853	14.51
3	17(144)	4	0.906	0.853	15.05
4	16(136)	4	0.904	0.851	15.16
5	15(124)	4	0.901	0.849	14.74
6[a]	14(116)	4	0.901	0.849	14.73
7	13(104)	4	0.897	0.847	15.16
8	12(96)	4	0.896	0.846	15.12
9	11(88)	4	0.900	0.843	16.61
10	10(80)	4	0.901	0.842	16.11
CoMFA model	672(5712)	4	0.892	0.681	35.14

[a] Represents the best protected chromosome.

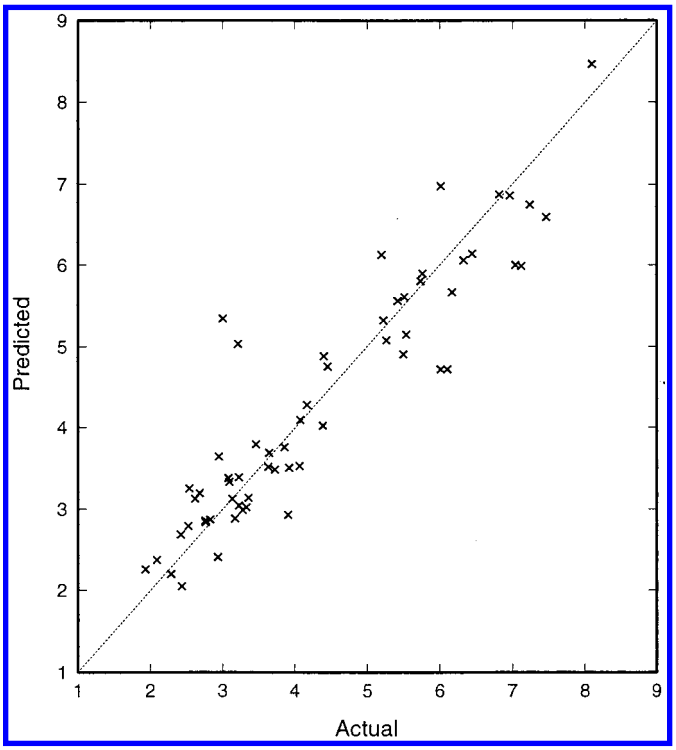


Figure 3. Plot of the observed and predicted inhibitory activities by the final GARGS model.

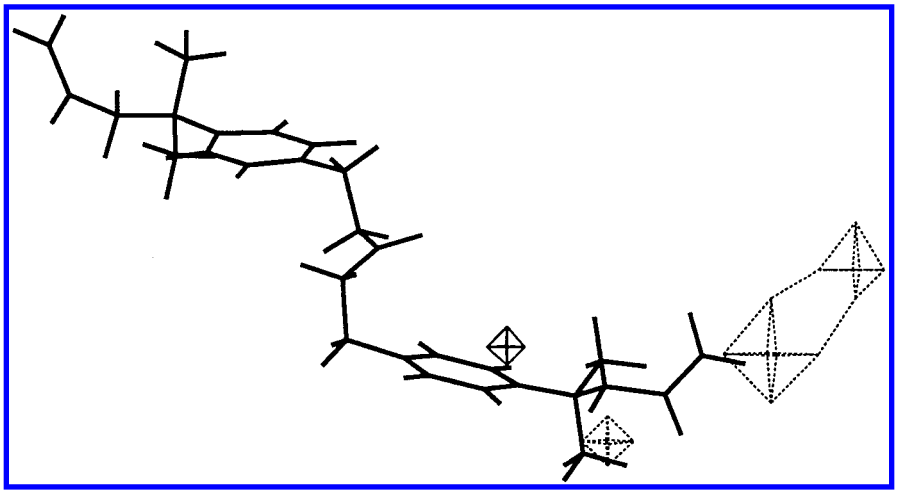


Figure 4. Steric contour map of the final GARGS model. Solid contour represents the sterically favorable region. Dotted contour represents the sterically unfavorable region. BW284C51 is displayed for reference.

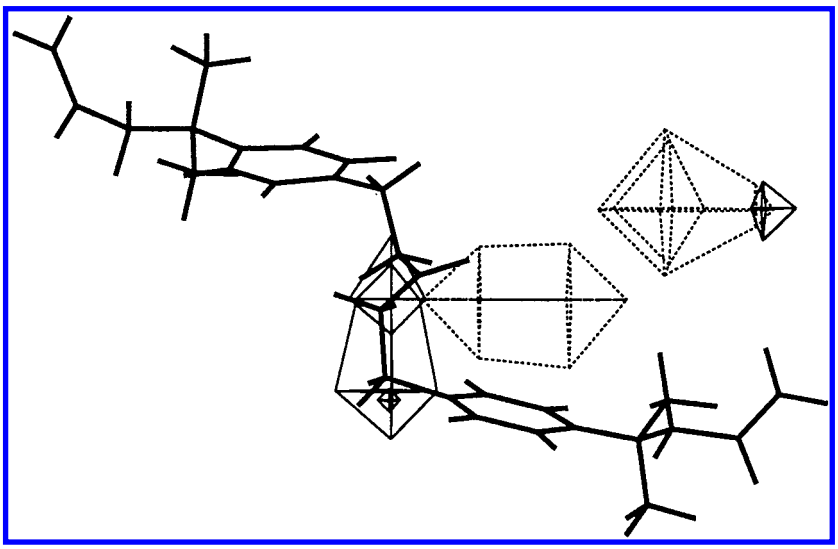


Figure 5. Electrostatic contour map of the final GARGS model. Solid contour represents the favorable region for the positively charged group. Dotted contour represents the favorable region for the negatively charged group. BW284C51 is displayed for reference.

model has an improved q^2 value, but its fitting value (the r^2 value) is worse compared with that of conventional CoMFA model (0.852 versus 0.892). This inferiority motivated an alternative GARGS study with a more refined resolution.

Finally, GARGS with 8-6-7 resolution was applied to the CoMFA field table. The values of GA parameters are summarized in Table 2. Table 4 shows the 10 high-ranked chromosomes with protection in the final GA population. The statistics of the conventional CoMFA model are also listed in Table 4 for reference. Table 4 indicates two minimum points of the PRESS value (chromosome nos. 2 and 6). Since two chromosomes have the very similar coefficient contour maps, chromosome 6 was selected as the best model judged from the minimum number of field variables. The plot of the observed versus predicted inhibitory activities is shown in Figure 3. Because both the internal and external predictions were significantly improved compared with that of the conventional CoMFA model (q^2: 0.681 versus 0.849, PRESS: 35.14 versus 14.73), this model (chromosome 6 in Table 4) was determined to be the final GARGS model in this study.

3.3. Comparison of the GARGS Model with the Active Site in AChE. To visualize the information content of the final GARGS model, the coefficient contour maps were generated. Figures 4 and 5 show the steric and electrostatic coefficient contour maps, respectively. All contour maps were drawn at the 0.015 level, and they showed the lattice points whose values were interpolated as 0.015 (positive) and −0.015 (negative), respectively. BW284C51 with the highest inhibitory activity was taken as the reference compound for specifying 3D space. The dotted contour in Figure 4 represents the region of unfavorable steric effect, while the solid contour represents the region of high steric tolerance. The small solid region is situated close to the benzene ring, and the introduction of some appropriate substituents to the benzene ring is favorable for activity. The presence of the dotted region near the end of ethylene chain indicates that activity is not be favored by a long chain. Also, the unfavorable region for steric interaction is observed below the ammonium ion center. The solid contour in Figure 5 describes the region where the positively charged groups enhance activity, and dotted contour describes the region where the negatively charged groups enhance activity. It should be noted that the two dotted contours are located above the molecule, while the small solid contour is close to the center of the molecule.

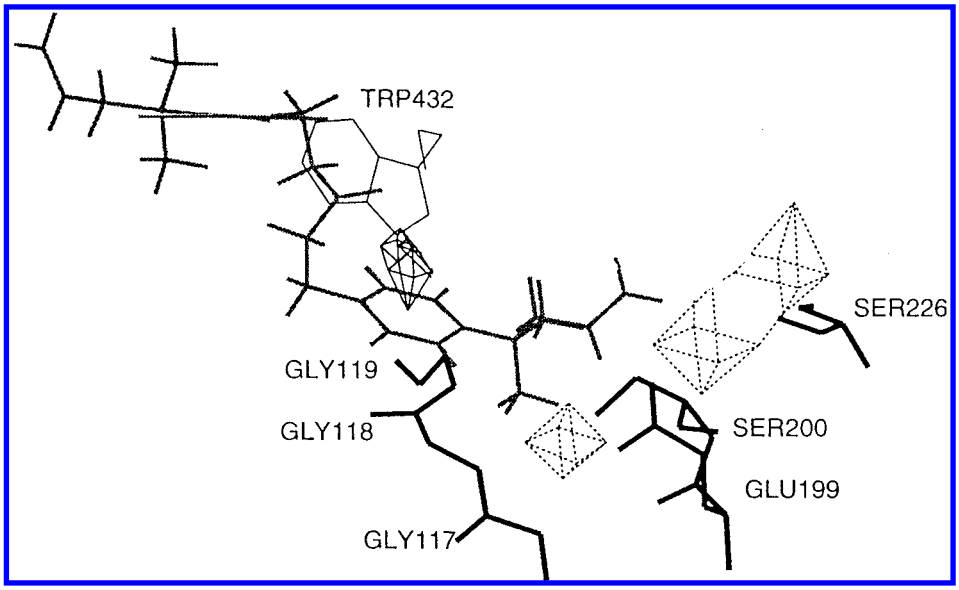


Figure 6. Steric contour map with the important active site residues in AChE.

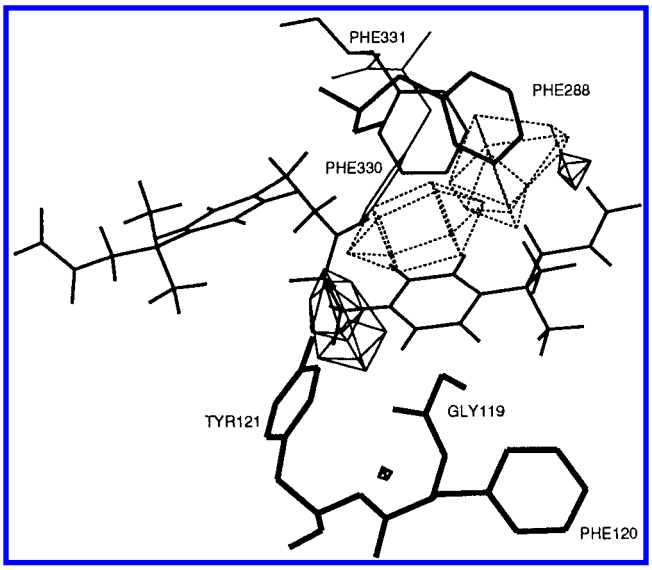


Figure 7. Electrostatic contour map with the important active site residues in AChE.

We are interested to see whether the results would be consistent with the steric and electrostatic properties of the active site in AChE. Thus, the coefficient contour maps and the active site were superimposed in 3D space and were compared with each other. The steric and electrostatic fields and the important active site residues are shown in Figures 6 and 7, respectively. The level of agreement was very high from graphical investigation. For the steric field, the solid region is located in a large cleft made from GLY119 and TRP432. The large dotted region contacts with SER200 and SER226, while the small dotted region is close to the GLY117-118 and GLU199-SER200. For the electrostatic field, the solid region is positioned to make favorable electrostatic interaction with the amide bond of GLY119 and the phenol portion of TYR121. The two dotted regions are surrounded by three hydrophobic side chains of PHE288, PHE330, and PHE331. This nice complementary relationships between the coefficient contour maps and the active site in AChE indicate the powerful ability of GARGS for reproducing a real and meaningful CoMFA model.

4. CONCLUSION

In the present paper, we applied GARGS to the data set of AChE inhibitors to evaluate whether GARGS can pinpoint known molecular interactions in 3D space. GARGS successfully gave the best model whose coefficient contour maps were consistent with the steric and electrostatic properties of the active site in AChE.

In recent years, structure-based drug design (SBDD)[29] has been a major paradigm for lead identification and optimization. SBDD is the rational design of molecules based on the 3D structure of the target receptor. This method is useful, but the accuracy to predict binding affinities is probably limited by the approximation used in the force field and electrostatic calculations. The greater computer power and deeper insight into the biophysics of the target receptor are required for high prediction. GARGS may facilitate the prediction of binding affinities if one already has a series of compounds and assumes the same binding mode. Moreover, multivariate 3D-QSAR analysis can also be done using the sophisticated probe atoms provided by the GRID software.[30] The multiple probes in GRID can describe ligand−receptor interactions in more detail. The amount of interaction field variables from GRID can be reduced and simplified by GARGS, and then the predictive 3D-QSAR model simulating the physical or chemical environment in receptor may be obtained.

ACKNOWLEDGMENT

We thank Dr. Sung Jin Cho at the University of North Carolina for kindly providing us the 3D coordinates of AChE inhibitors. We also thank Prof. Zdenek Slanina at the Toyohashi University of Technology and a reviewer for help in improving the English.

REFERENCES AND NOTES

(1) Green, S. M.; Marshall, G. R. 3D-QSAR: A Current Perspective. *Trends Pharmcol. Sci.* **1995**, *16*, 285−291.

(2) 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.

(3) Geladi, P.; Kowalski, B. R. Partial Least Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(4) van de Waterbeemd, H. *Chemometric Methods in Molecular Design, Methods and Principles in Medicinal Chemistry Vol 2*; Verlag Chemie: Weinheim, 1995.
(5) Kubinyi, H. *3D QSAR in Drug Design, Theory and Applications*; ESCOM: Leiden, 1993.
(6) SYBYL version 6.2; Tripos Inc.: St. Louis, MO 63144.
(7) Dean, P. M. *Molecular Similarity in Drug Design*; Blackie Academic: Glasgow, 1995.
(8) Klebe, G.; Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70−80.
(9) Sanz, F.; Giraldo, J.; Manaut, F. *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Applications*; Prous Science: New York, 1995.
(10) Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quantum Struct.-Act. Relat.* **1993**, *12*, 137−145.
(11) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of Comparative Molecular Field Analysis Models: Effects of Data Scaling and Variable Selection Using a Set of Human Synovial Fluid Phospolipase A_2 Inhibitors. *J. Med. Chem.* **1997**, *40*, 1136−1148.
(12) Lingren, F.; Geladi, P.; Rannar, S.; Wold, S. Interactive Variable Selection (IVS) for PLS Part I: Theory and Algorithm. *J. Chemom.* **1994**, *8*, 349−363.
(13) 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. *Quantum Struct.-Act. Relat.* **1993**, *12*, 9−20.
(14) Cho, J. S.; 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.
(15) 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.
(16) 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)* **1998**, *425*, 255−262.
(17) Kimura, T.; Hasegawa, K.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 276−282.
(18) 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.
(19) Norinder, U. Single and Domain Mode Variable Selection in 3D QSAR Applications. *J. Chemom.* **1996**, *10*, 95−105.
(20) Cho, S. J.; Garsia, M. L. S.; Bier, J.; Tropsha, A. Structure-Based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.* **1996**, *39*, 5064−5071.
(21) 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.
(22) Dunn, W. J.; Rogers, D. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: San Diego, 1996; pp 109−130.
(23) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quantum Struct.-Act. Relat.* **1994**, *13*, 285−294.
(24) Goldberg, D. E. *Genetic Algorithm in Search, Optimization and Machine Learning*; Addison-Wesley: New York, 1989.
(25) Roger, D. E. *Procedural Elements for Computer Graphics*; McGraw-Hill: New York, 1985.
(26) Davis, L. *Handbook of Genetic Algorithm*; Van Nostrand Reinhold: New York, 1991.
(27) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.
(28) Leardi, R. Application of Genetic Algorithm to Feature Selection under Full Validation Conditions and to Outliner Detection. *J. Chemom.* **1994**, *8*, 65−79.
(29) Verlinde, C. L.; Hol, W. G. Structure-based drug design: progress, results and challenges. *Structure* **1994**, *2*, 577−587.
(30) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

CI980088O