Enhancement of Ordinal CoMFA by Ridge Logistic Partial Least Squares

Takanori Ohgaru,[†,‡] Ryo Shimizu,[§] Kosuke Okamoto,[†] Norihito Kawashita,[†,||] Masaya Kawase,[⊥]
Yuko Shirakuni,[†] Rika Nishikiori,[⊥] and Tatsuya Takagi*[,†,||]

Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871, Japan, Medicinal Chemistry Laboratory, Mitsubishi Tanabe Pharma Corporation, 3-16-89 Kashima, Yodogawa-Ku, Osaka 532-8505, Japan, Corporate Strategy Department, Mitsubishi Tanabe Pharma Corporation, 3-2-10, Dosho-machi, Chuo-Ku, Osaka 541-8505, Japan, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan, and Faculty of Pharmacy, Osaka Ohtani University, 3-11-1 Nishikiorikita, Tondabayashi, Osaka 584-8540 Japan

Received November 29, 2007

Conventional comparative molecular field analysis (CoMFA) requires at least 3 orders of experimental data, such as IC_{50} and K_i, to obtain a good model, although practically there are many screening assays where biological activity is measured only by rating scale. To improve three-dimensional quantitative structure–activity relationship (3D-QSAR) analysis, we developed in this study a modified ordinal classification-oriented CoMFA using partial-least-squares generalized linear regression and ridge estimation. The modified Logistic CoMFA was validated using a corticosteroid binding globulin receptor binding data set, a benchmark for 3D-QSAR, and an acetylcholine esterase inhibitor data set. Our results show that modification of Logistic CoMFA enhanced both prediction accuracy and 3D graphical analysis. In addition, the 3D graphical analysis of the modified Logistic CoMFA was much improved. This improvement resulted in more accurate information on the binding mode between proteins and ligands than in the case of conventional CoMFA.

INTRODUCTION

Quantitative structure–activity relationships (QSAR) are used to establish a correlation between chemical structure and specific biological activity,[1] and the derived models are used to predict the activity of untested compounds. This correlation is one of the most important steps in drug discovery, particularly in the hit-to-lead stage. The prevalence of commercial cheminformatics tools, such as Sybyl, makes it convenient to perform three-dimensional QSAR (3D-QSAR) analysis. Above all, comparative molecular field analysis (CoMFA)[2] is a widely used approach for generating descriptors based on 3D structural information of molecules.

In real screening, many assays measure compounds' biological activity only by rating scale. Under such circumstances, it is difficult to obtain good CoMFA models, since relatively accurate experimental IC_{50} and pK_i values are generally required. In a preceding paper, we have proposed an ordinal classification approach using CoMFA (Logistic CoMFA) and showed that this approach is better and more robust than conventional CoMFA with rating scale activity.[3]

Logistic CoMFA couples CoMFA with ordinal logistic regression (OLR), which classifies samples according to the probability of each rank. Unfortunately, ordinary algorithms of logistic regression analysis do not converge in some cases.[4,5] Infinite parameter estimates can occur depending on the configuration of the sample points in the observation space.[6]

Using PLS with penalized logistic regression, Fort and Lacroix proposed a robust samples classification.[7] Their method is based on ridge estimators in the logistic regression reported by Le Cessie and Van Houwelingen.[8] Both methods, however, were developed for analyzing a binary response variable, and there is no way of applying them to ordinal classification. Although multigroup iteratively reweighted partial least squares (MIRWPLS) was generalized for multigroup classification by Ding and Gentleman,[9] this method is in principle used to analyze nominal data. Thus, MIRWPLS treats the ratings as nominals with no special ordering. Additionally, with MIRWPLS, it is difficult to estimate parameters because, unlike MIRWPLS for binary data analysis, MIRWPLS uses the (number of classes $-$ 1)^2-fold size of a sparse matrix for explanatory variables. By and large, CoMFA uses as many as >1 000 explanatory variables and requires a huge-sized matrix of variables. Hence, applying MIRWPLS to CoMFA is impracticable.

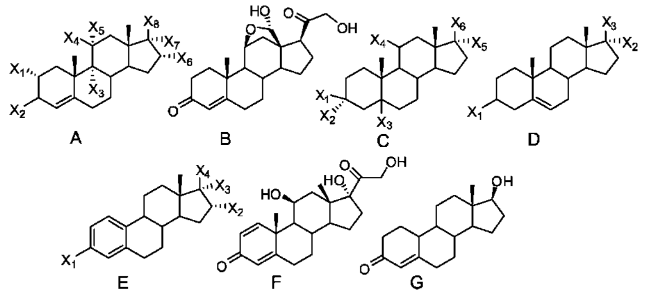
Recently, Bastein et al.[10] contrived a new logistic PLS algorithm, which is based on the PLS generalized linear regression (PLS-GLR) model. The PLS-GLR approach performs OLR analysis on every explanatory variable without enlarging the size of the explanatory matrix in the process of computation of the latent variables.

In this study, we modified Logistic CoMFA by harnessing two approaches. The first modification is the application of Logistic CoMFA to PLS-GLR instead of ordinary OLR-based PLS. The second is the incorporation of Logistic CoMFA with ridge penalty estimation, although, as mentioned above, the occurrence of a convergence problem was a possibility in some cases. Next, we compared the modified ordinal classification CoMFA with Logistic CoMFA.

* Corresponding author e-mail: satan@gen-info.osaka-u.ac.jp.
[†] Graduate School of Pharmaceutical Sciences, Osaka University.
[‡] Medicinal Chemistry Laboratory, Mitsubishi Tanabe Pharma Corporation.
[§] Corporate Strategy Department, Mitsubishi Tanabe Pharma Corporation.
[||] Research Institute for Microbial Diseases, Osaka University.
[⊥] Faculty of Pharmacy, Osaka Ohtani University.

10.1021/ci700444z CCC: $40.75 © 2008 American Chemical Society
Published on Web 03/14/2008

Table 1. Chemical Structures and CBG Activity of Steroids



	steroids	pK_i	class	MCS	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
training	cortisol	7.88	1	A	H	=O	H	OH	H	H	OH	$COCH_2OH$
training	corticosterone	7.88	1	A	H	=O	H	OH	H	H	H	$COCH_2OH$
training	11-deoxycortisol	7.88	1	A	H	=O	H	H	H	H	OH	$COCH_2OH$
training	17α-hydroxyprogesterone	7.74	1	A	H	=O	H	H	H	H	OH	COMe
training	11-deoxycorticosterone	7.65	1	A	H	=O	H	H	H	H	H	$COCH_2OH$
training	progesterone	7.38	1	A	H	=O	H	H	H	H	H	COMe
training	cortisone	6.89	2	A	H	=O	H	H	H	H	OH	$COCH_2OH$
training	testosterone	6.72	2	A	H	=O	H	H	H	H	H	OH
training	aldosterone	6.28	2	B								
training	dihydrotestosterone	5.92	2	C	H	=O	H	H	H	OH		
training	4-androstenedione	5.76	3	A	H	=O	H	H	H	H		=O
training	androsterone	5.61	3	C	OH	H	H	H		=O		
training	etiocholanolone	5.23	3	D	OH		=O					
training	pregnenolone	5.23	3	D	OH		COMe					
training	androstanediol	5.00	3	C	H	OH	H	H	H	OH		
training	5-androstenediol	5.00	3	D	OH		OH					
training	dehydroepiandrosterone	5.00	3	D	OH		=O					
training	estradiol	5.00	3	E	OH	H	H	OH				
training	estriol	5.00	3	E	OH	OH	H	OH				
training	estrone	5.00	3	E	OH	H		=O				
training	17α-hydroxypregnenolone	5.00	3	D	OH	OH	COMe					
test	2α-methylcortisol	7.69	1	A	Me	=O	H	OH	H	H	OH	$COCH_2OH$
test	cortisol acetate	7.55	1	A	H	=O	H	OH	H	H	OH	$COCH_2OMe$
test	predonisolone	7.51	1	F								
test	epicorticosterone	7.20	1	A	H	=O	H	H	OH	H	H	$COCH_2OH$
test	16α-methylprogesterone	7.12	1	A	H	=O	H	H	H	Me	H	COMe
test	19-norprogesterone	6.82	2	A	H	=O	H	H	H	H	H	COMe
test	4-pregnene-3,11,20-trione	6.78	2	A	H	=O	H	=O		H	H	COMe
test	16α,17α-dihydroxyprogestero	6.25	2	A	H	=O	H	H	H	OH	OH	COMe
test	19-nortestosterone	6.14	2	G								
test	2α-methyl-9α-fluorocortisol	5.80	2	A	Me	=O	F	OH	H	H	OH	$COCH_2OH$

Hence, we present here a new approach for the improvement and applicability of modified Logistic CoMFA with PLS-GLR and ridge estimation and a comparison of the improved Logistic CoMFA with conventional CoMFA and original Logistic CoMFA using two data sets. One of these data sets is the corticosteroid binding globulin (CBG) receptor binding data set,[11] which is widely used as a benchmark for 3D-QSAR. The other data set is the acetylcholine esterase (AChE) inhibitor data set. Each data set was analyzed from two aspects of the CoMFA method. The first aspect is prediction of the accuracy of the modified Logistic CoMFA method. Accurate prediction of Logistic CoMFA enables us to effectively prioritize the screening of certain compounds. The second is contour map analysis, which can identify important portions for interaction between the protein and ligand.

METHODS

1. Data Set. *1.1. CBG Data Set.* Activities and chemical structures of CBG data set ligands were used for validation of the modified Logistic CoMFA and for comparison of this new method with original Logistic CoMFA. Some groups have reported validation of each 3D-QSAR using the CBG data set as a benchmark.[12–15] The CBG data set comprises 21 compounds for training and 10 compounds for testing. Not only activity values but also rating classes can be obtained.[16]Although there is no steroid of $class_3$ in the test set, it is desirable that all rating classes be included in the test set. However, we used the data set without any modification, such as shuffling between training and test sets, because the CBG data set is a 3D-QSAR benchmark (Table 1).

1.2. AChE Data Set. Activities and chemical structures of AChE data set ligands were also used for validation of the modified Logistic CoMFA. The AChE data set consists of a series of 111 inhibitors (74 training and 37 test compounds). We used in this study 3D coordinates and partial charges reported by Sutherland et al.[17] pIC_{50} values ranged widely from 4.27 to 9.52, and activity classes were allocated as follows: $class_1$ ($pIC_{50} \geq 7.5$), $class_2$ ($6.0 \leq pIC_{50} < 7.5$), and $class_3$ ($pIC_{50} < 6.0$) (Figure 1). Unlike the CBG data set, the test set of the AChE data set comprised all activity classes.

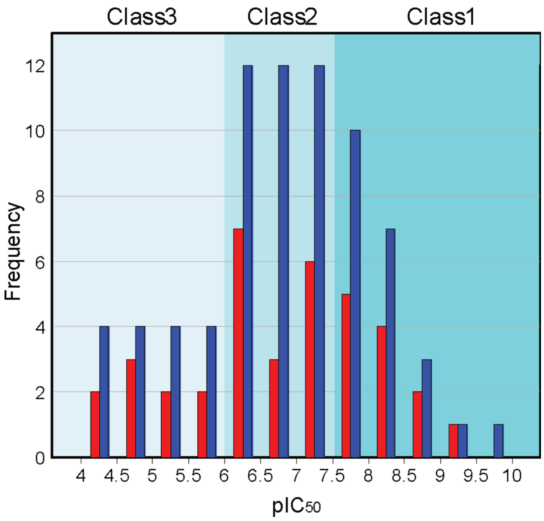


Figure 1. Distribution of inhibitory activity for AChE training (blue) and test (red) sets.

2. Molecular Modeling. The 3D coordinates were used without any refinement for both data sets. For the CBG data set, Gasteiger−Hückel charges were assigned to each atom by Sybyl, version 7.22 (Tripos Inc.). For the AChE data set, atomic partial charges in mol2 files obtained from Sutherland et al.'s study[17] were used.

3. Calculation of Steric and Electrostatic Potential Fields. The steric and electrostatic CoMFA potential fields were calculated at each lattice intersection of a regularly spaced grid as implemented in Sybyl using Lennard-Jones and Coulomb potentials, respectively. Calculations were performed with Sybyl standard parameters.

4. Calculation of Latent Variables in Ordinal CoMFA, Based on the PLS-GLR Model. The calculation of latent variables was based on the PLS-GLR method,[10] where latent variables were extended from linear models to generalized linear models.

Algorithm of Ordinal Logistic PLS

Computation of the first latent variable $\mathbf{t}_1$:

1-1. OLR of **y** was performed on each $\mathbf{x}_j$ (j = 1, 2, . . ., p) and logistic regression coefficient a_{1j}. a_{1j} were regarded as 0 when the significance level of a_{1j} was low (p value > 0.25).

1-2. $\mathbf{w}_1$, the weight column vector toward **X**, was calculated as $\mathbf{w}_1 = \mathbf{a}_1/|\mathbf{a}_1|$, where $\mathbf{a}_1 = (a_{11}, a_{12}, ..., a_1p)$.

1-3. The first latent variable $\mathbf{t}_1$ was determined as $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, . . ., \mathbf{x}_p)$.

Computation of the second latent variable $\mathbf{t}_2$:

2-1. A simple linear regression of each $\mathbf{x}_j$ on $\mathbf{t}_1$ and calculation of each residual $\mathbf{x}_{2j}$ were performed.

2-2. OLR of **y** on $\mathbf{t}_1$ and each $\mathbf{x}_{2j}$ and calculation of logistic regression coefficient a_{2j} on $\mathbf{x}_{2j}$were performed. The forward variables selection method was used. a_{2j} was also regarded as 0 when the significance level of a_{2j} was low (p value > 0.25).

2-3. $\mathbf{w}_2$ was calculated as $\mathbf{w}_2 = \mathbf{a}_2/|\mathbf{a}_2|$, where $\mathbf{a}_2 = (a_{21}, a_{22}, . . ., a_{2p})$.

2-4. $\mathbf{t}_2$ was determined as $\mathbf{t}_2 = \mathbf{X}_2\mathbf{w}_2$, where $\mathbf{X}_2 = (\mathbf{x}_{21}, \mathbf{x}_{22}, . . ., \mathbf{x}_{2p})$.

Computation of the ith latent variable $\mathbf{t}_i$ ($i \geq 3$):

3-1. Multilinear regression of each $\mathbf{x}_{i-1j}$ and $\mathbf{t}_1, \mathbf{t}_2, . . ., \mathbf{t}_{i-1}$ and calculation of each residual $\mathbf{x}_{ij}$ were preformed.

3-2. OLR of **y** on $\mathbf{t}_1, \mathbf{t}_2, . . ., \mathbf{t}_{i-1}$ and each $\mathbf{x}_{ij}$ and calculation of logistic regression coefficient a_{ij} on $\mathbf{x}_{ij}$ were performed. The forward variables selection method was used. a_{ij} was regarded as 0 when the significance level of a_{ij} was low (p value >0.25).

3-3. $\mathbf{w}_i$ was calculated as $\mathbf{w}_i = \mathbf{a}_i/|\mathbf{a}_i|$, where $\mathbf{a}_i = (a_{i1}, a_{i2}, . . ., a_{ip})$.

3-4. $\mathbf{t}_i$ was determined as $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i$, where $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, . . ., \mathbf{x}_{ip})$.

Rule to stop computing latent variables:

Latent variables were successively computed as shown above. If $\mathbf{a}_h$ was equivalent to **0** in the computation of the hth latent variable, computation was terminated.

Although the original PLS-GLR algorithm normalizes each residual before calculation of a new latent variable, PLS-GLR normalization of residuals was not adopted in this study.

5. Modification to Ordinal CoMFA by Ridge OLR. Latent variables calculated as shown above were applied to explanatory variables of OLR analysis. The followings are details of the modification to Ordinal CoMFA.

OLR analysis gives the probability of each rank. For instance, the data set is categorized into three rating classes as follows:

$$\mathrm{Prb}(\mathrm{Class}_1) = \{1 + \exp(-\eta_1)\}^{-1}$$

$$\mathrm{Prb}(\mathrm{Class}_2) = \{1 + \exp(-\eta_2)\}^{-1} - \{1 + \exp(-\eta_1)\}^{-1}$$

$$\mathrm{Prb}(\mathrm{Class}_3) = 1 - \{1 + \exp(-\eta_2)\}^{-1}$$

where η_1 and η_2 can be rewritten as

$$\eta_1 = c_1 - \mathbf{b't}$$

$$\eta_2 = c_2 - \mathbf{b't}$$

$$c_1 \leq c_2$$

The coefficients c_1, c_2, and **b** (size $q \times 1$) were evaluated using a maximum likelihood estimation (MLE) with a ridge penalty. With Prb_{ij}, the probability when the activity rating of compound i (i = 1, 2, . . ., n) ranks class j (j = 1, 2, . . ., m), the likelihood (L) is obtained by

$$L = \prod_{i=1}^{n} \mathrm{Prb}_{i1}^{y_1} \times \mathrm{Prb}_{i2}^{y_2} \dots \mathrm{Prb}_{im}^{y_m}$$

where n is the number of compounds, y_j (j = 1, 2, 3) is 0 or 1, and $y_1 + y_2 + \cdots + y_m = 1$.

In MLE, the coefficients c_1, c_2, and **b** in the case where L is a maximum are determined as the most appropriate coefficients. The log-likelihood is good to estimate unknown parameters with l set as ln$L$

$$l = \sum_{i=1}^{n} (y_1 \ln \mathrm{Prb}_{i1} + y_2 \ln \mathrm{Prb}_{i2} + \dots + y_m \ln \mathrm{Prb}_{im})$$

Unfortunately, logistic regression sometimes encounters obstacles, such as a complete or quasi-complete separation problem.[6] To avert such a problem, we adopted the penalty approach in Fisher's score method[18] and the information criteria method.[19] $n|\mathbf{b}|^2$ was introduced as a penalty term.

$$g(c_1,c_2,\mathbf{b}) = l - \lambda n|\mathbf{b}|^2$$

where λ is the weight of the penalty term ($\lambda \geq 0$). A conjugate-gradient numerical optimization algorithm was

Table 2. Summary of Leave-One-Out Cross-Validation of CBG Training Set of the Modified Logistic and Original Logistic CoMFA Analyses

	CoMFA	
	modified	logistic
q_s[a]	0.79	0.75
no. of correct	14	13
accuracy	67%	62%

[a] Cross-validated Spearman's rank correlation coefficient.

Table 3. Prediction of CBG Test Set by the Modified Logistic and Original Logistic CoMFA Analyses

	CoMFA	
	modified	logistic
q_s[a]	0.45	0.45
no. of correct	7	7
accuracy	70%	70%

[a] Cross-validated Spearman's rank correlation coefficient.

adopted to maximize function g. In this study, 0, 0.0001, 0.001, 0.01, and 0.1 were used as λ. To estimate the coefficients of OLR in any case, the penalty approach was used.

Details of PLS-GLR with ridge estimation by are described in the Appendix.

6. Difference between the Precedent Logistic CoMFA and the Modified CoMFA. In the precedent Logistic CoMFA algorithm, latent variables are calculated to maximize the correlation between cumulative probabilities of activity ratings and the respective latent variable. Strictly speaking, this calculation is a kind of ordinary PLS. Latent variables calculated in the process of PLS are used for OLR.

On the other hand, in the modified Logistic CoMFA algorithm, latent variables are calculated in parallel with OLR. Thus, the latent variables used are appropriate for OLR.

Although the difference is small in the calculation of the first latent variable, with an increase in the number of latent variables to calculate, the difference becomes remarkable.

7. CoMFA Program. Both the modified Logistic CoMFA and the original Logistic CoMFA, programmed in Fortran90, were computed on a dual-core Xeon 2.0 GHz computer.

RESULTS AND DISCUSSION

1. Steroids in CBG Binding Analysis. *1.1. Validation of CBG Data Set.* The modified Logistic CoMFA and the original Logistic CoMFA models were performed with leave-one-out (LOO) cross-validation (Table 2). The results of the modified Logistic CoMFA were obtained with $\lambda = 0$. This means no penalty was added. Both models were good in terms of the cross-validated Spearman's rank correlation coefficient (q_s), but the modified Logistic CoMFA was found to be slightly better for accurate activity ratings. The prediction capability of each model was next investigated using the best numbers of latent variables. As a result, the modified Logistic CoMFA was found to be as accurate as the original Logistic CoMFA (Table 3). This is probably because the original Logistic CoMFA is already established as a prediction model for activity ratings. Therefore, modification of Logistic CoMFA had no improving effect on prediction accuracy.

1.2. Contour Interpretation. The contour maps obtained by CoMFA show how 3D-QSAR methods are useful to identify features for recognizing protein–ligand interactions. CoMFA steric interactions are represented by favored green and disfavored yellow contours, while electrostatic interactions are represented by negative-charge favored red and positive-charge favored blue contours. Figure 2 shows a comparison of contour maps (standard deviation × coefficient) derived from (a and b) the modified Logistic CoMFA, (c) conventional CoMFA, and (d) the original Logistic CoMFA. A reduction of C-3 causes low binding affinity, which is supported by all contour maps. The conventional CoMFA map supports the fact that the carbonyl at the 17 position causes low activity, though the original Logistic CoMFA map, unfortunately, does not support such a fact. On the other hand, the modified Logistic CoMFA map supports the importance of the carbonyl at the 17 position. These findings indicate that important information is retained by the use of ordinal classes, although conventional CoMFA builds a 3D-QSAR model using pK_i values, while both original and modified Logistic CoMFA use pK_i classes that are rounded pK_i's.

The modified Logistic CoMFA map was found to be as good as the conventional CoMFA map. Furthermore, to validate the performance of the modified Logistic CoMFA, we compared the CoMFA contour map with X-ray data of the binding between CBG and cortisol.[20] Figure 3 shows a closeup of cortisol in the steroid binding site of CBG. The electrostatic potential on the CBG surface is obviously distributed, and it is clear that the cortisol binds to CBG complementarily. Of the five polar atoms in cortisol, only two oxygen atoms connected with the carbon atoms C-11 and C-20 of the steroid interact directly with CBG residues.[20] The hydroxyl oxygen atoms at C-11 and C-20 make hydrogen bonds to Asp256 and Gln224, respectively (Figure 4). Moreover, Figure 3 holds the fact that there is a broad space around C-17 of the cortisol. This is supposed by the CoMFA contour map. Especially, C-3 and C-17 of the cortisol and their respective surrounding on the CBG surface indicate that these steric as well as electrostatic interactions are important for the binding.

2. AChE Binding Analysis. *2.1. Validation of AChE Data Set.* Shown in Table 4 are the results from the LOO cross-validation of each CoMFA method. The results of the modified Logistic CoMFA were obtained with $\lambda = 0.0001$. (Under $\lambda = 0$, the modified Logistic CoMFA encountered a separation problem as mentioned above.) Both models were good in terms of q_s, but the modified Logistic CoMFA was found to be slightly better for accurate activity ratings. Next, the prediction capability of each model was investigated using the best numbers of latent variables. As a result, the modified Logistic CoMFA was found to be as accurate as the original Logistic CoMFA (Table 5). This finding is consistent with that of CBG, indicating that the original Logistic CoMFA without any modification is a good 3D-QSAR model to predict activity ratings of untested compounds.

2.2. Contour Interpretation. Contour maps of the modified Logistic CoMFA and the conventional CoMFA are shown in Figure 5. E2020, marketed as Aricept, is depicted in the center. Both maps show that steric potential is as important

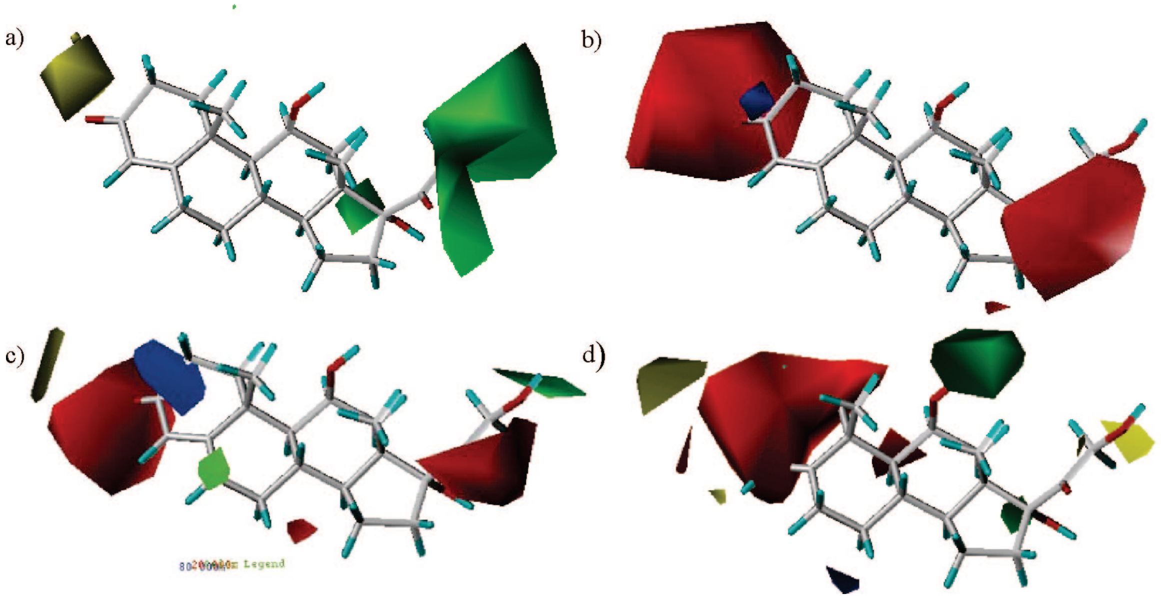


Figure 2. CBG contour maps (standard deviation × coefficient) derived from the modified Logistic CoMFA, conventional CoMFA, and original Logistic CoMFA: (a) steric effect by the modified Logistic CoMFA; (b) electrostatic effect by the modified Logistic CoMFA; (c) steric and electrostatic effects by conventional CoMFA; (d) steric and electrostatic effects by original Logistic CoMFA. The centered molecule, cortisol, is strongly bound to CBG. To increase the activity, the positive charge (in blue) and the negative charge (in red) have to be increased. In addition, molecular volume has to be increased (green) or decreased (yellow) to increase the activity.

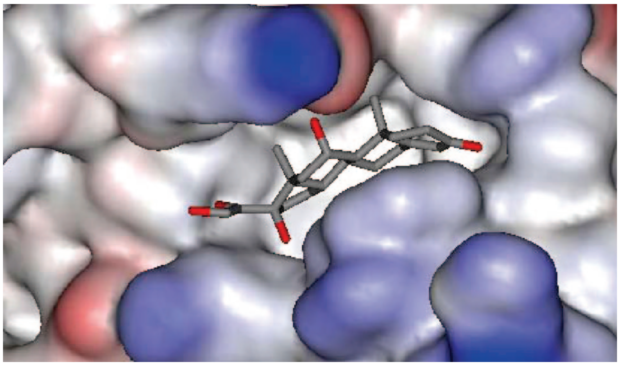


Figure 3. Closeup of cortisol in the steroid binding site of CBG (X-ray data, 2v95, from the Protein Data Bank). The CBG protein surface is colored according to molecular electrostatic potential (positive charge in blue and negative charge in red).

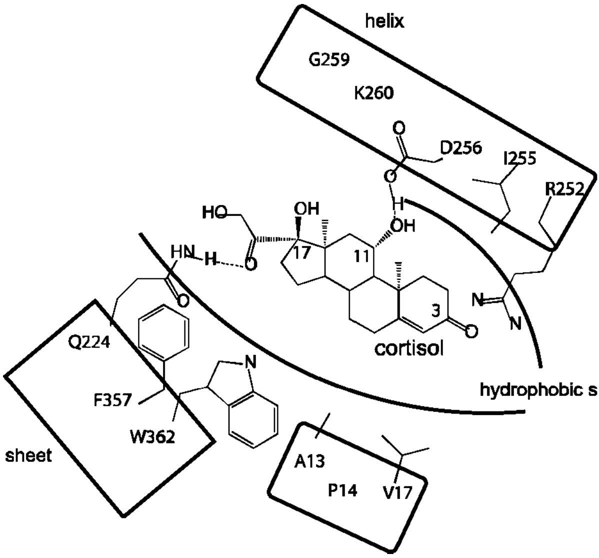


Figure 4. Schematic depiction of the interactions between CBG and cortisol. Hydrogen bonds are shown with dotted lines (X-ray data, 2v95, from the Protein Data Bank).

as electrostatic potential for inhibitory activity. The modified Logistic CoMFA map clearly shows that there is a narrow space near the carbonyl group of 1-indanon and that there is a broad space around the positions C-3 and C-4 of 1-indanon and another space nearby the dimethoxy group at C-6. The conventional CoMFA map only shows a broad space at the opposite of the carbonyl group of 1-indanon.

In order to validate both CoMFA maps, we compared contour maps with X-ray data of the binding between AChE and E2020.[21] Figure 6 shows a close up of E2020 in the binding site of AChE. The electrostatic potential on the AChE surface is not clearly localized, indicating that interactions other than electrostatic interaction are important for the binding. There is indeed no polar interaction between AChE and E2020 (Figure 7). As for steric interaction, X-ray data show a broad space around positions C-3 and C-4 of 1-indanon, a space near the dimethoxy group at C-6, and a narrow space near the carbonyl group. E2020 binds along the active site and interacts with the peripheral anionic site.[21] Thus, the modified Logistic CoMFA map gives more accurate information than the conventional CoMFA or the ordinal classification-based CoMFA (Ordinal CoMFA) and facilitates understanding of the structure–activity relationship.

Table 4. Summary of Leave-One-Out Cross-Validation of AChE Training Set of the Modified Logistic and Original Logistic CoMFA Analyses

	CoMFA	
	modified	logistic
$q_s{}^a$	0.66	0.53
no. of correct	54	52
accuracy	73%	70%

[a] Cross-validated Spearman's rank correlation coefficient.

Using CBG and AChE data sets, we have found in this study that original Logistic CoMFA is a good model to predict activity ratings of untested compounds, but not a good

Table 5. Prediction of AChE Test Set by the Modified Logistic and Original Logistic CoMFA Analyses

	CoMFA	
	modified	logistic
q_s[a]	0.66	0.66
no. of correct	25	25
accuracy	68%	68%

[a] Cross-validated Spearman's rank correlation coefficient.

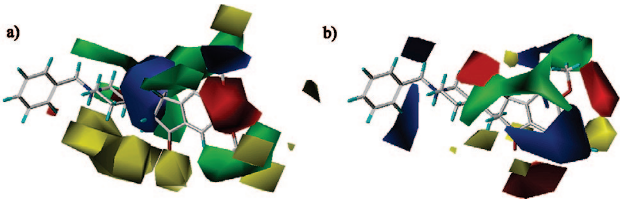


Figure 5. AChE contour maps (standard deviation × coefficient) derived from the modified Logistic CoMFA (a) and conventional CoMFA (b). The centered molecule, E2020, is strongly bound to AChE. To increase the activity, the positive charge (in blue) and the negative charge (in red) have to be increased. In addition, molecular volume has to be increased (green) or decreased (yellow) to increase the activity.

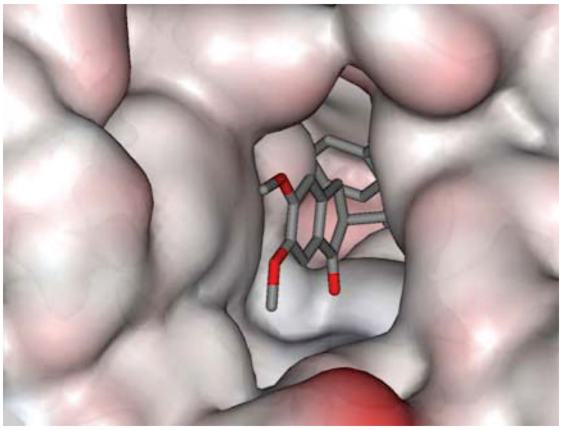


Figure 6. Close up of E2020 in the binding site of AChE (X-ray data, leve, from the Protein Data Bank). AChE protein surface is colored according to molecular electrostatic potential (positive charge in blue and negative charge in red).

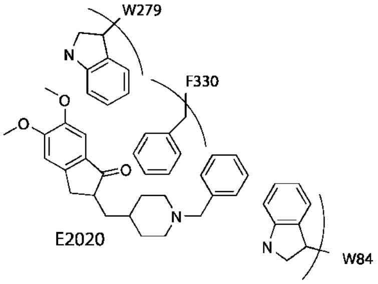


Figure 7. Schematic depiction of the interactions between AChE and E2020 (X-ray data, leve, from the Protein Data Bank).

model to build contour map analysis. In addition, our results show that modified Logistic CoMFA, which couples PLS-GLR and ridge estimation, is a good model for building contour map analysis. Furthermore, the modified Logistic CoMFA was found to exhibit a contour map as accurately as, or more accurately than, conventional CoMFA. The main reason for the excellence of Ordinal CoMFA with PLS-GLR and ridge estimation is probably overfitting. Conventional CoMFA uses PLS and produces several (normally at least > 3) latent variables. To avoid excessive fitting, a cross-validation method is used to determine the number of latent variables, which can influence the CoMFA contour map.[22] In contrast, Ordinal CoMFA uses logistic PLS, which also produces several (normally as much as < 3) latent variables. Using latent variables, Ordinal CoMFA estimates the probability of each rank and does not estimate activity ratings directly. Therefore, the number of latent variables used is not so influential.

CONCLUSION

In the present study, we modified Logistic CoMFA by incorporating PLS-GLR and ridge estimation into it and compared the modified Logistic CoMFA with original Logistic CoMFA and conventional CoMFA using CBG and AChE data sets. Our results show that modified Logistic CoMFA is superior in terms of both prediction accuracy and contour map analysis to other models. Especially, the modified Logistic CoMFA showed enhanced building of the contour map model and was as good as, or better than, conventional CoMFA. As lots of rank-scale biological activity data are produced in the process of drug discovery, we believe that Ordinal CoMFA is an important and powerful method to analyze rating data and to facilitate novel drug development.

NOMENCLATURE

The following notation is used. Uppercase bold variables are matrices, lowercase bold variables are column vectors, and lowercase variables are scalars.

$\mathbf{X}$: explanatory variables matrix (size $n \times p$)
$\mathbf{y}$: class probability variables vector (size $n \times 1$)
$\mathbf{x}_j$: jth explanatory variable vector (size $n \times 1$)
$\mathbf{w}_k$: column vector of logistic PLS weights matrix (size $p \times 1$)
$\mathbf{t}_k$: latent variable vector, a column vector of PLS scores matrix (size $n \times 1$)
$\mathbf{a}_h$: hth column vector of coefficients for OLR on elements of $\mathbf{x}_j$ (size $p \times 1$)
n: number of samples
p: number of X variables
j: integer counter for X variables
h: integer counter for latent variable dimension

ACKNOWLEDGMENT

We gratefully thank Dr. J. J. Sutherland for providing the AChE data set. We also wish to thank OpenEye Scientific Software, Inc. for allowing us to use the OEChem Toolkit free of charge. Finally, we are grateful to Drs. Ohmizu, Ogiku, Sunada, Kuroda, and Asao (Mitsubishi Tanabe Pharma Corporation) for their helpful advice and comments on QSAR.

APPENDIX

The procedure of ridge logistic PLS used in this study is now applied to a monoamine oxidase (MAO) inhibitor data set[23] as a case study. Table 6 shows normalized physical properties and inhibitory activity classes of the MAO data set. MAO inhibitory activities are categorized to four classes.

Table 6. Normalized Physical Properties and Inhibitory Activity Classes of MAO Data Set

ID	π^*	E_s^*	I^*	X^*	activity[a]
cpd_01	−0.69	0.49	1.95	−0.64	1
cpd_02	−0.79	1.08	−0.49	1.49	1
cpd_03	−0.69	1.08	−0.49	−0.64	1
cpd_04	0.14	0.36	−0.49	−0.64	1
cpd_05	−0.32	0.49	−0.49	−0.64	2
cpd_06	−0.32	0.49	−0.49	−0.64	2
cpd_07	−1.16	1.08	1.95	−0.64	2
cpd_08	−0.32	0.49	−0.49	1.49	3
cpd_09	−0.32	−0.72	−0.49	−0.64	3
cpd_10	0.61	−0.72	−0.49	−0.64	3
cpd_11	2.01	1.74	−0.49	−0.64	3
cpd_12	1.35	−0.76	−0.49	1.49	3
cpd_13	−0.97	−0.72	−0.49	−0.64	3
cpd_14	−0.32	0.49	−0.49	1.49	4
cpd_15	0.52	−1.49	−0.49	1.49	4
cpd_16	0.52	−1.49	−0.49	−0.64	4
cpd_17	0.05	−1.49	1.95	−0.64	4
cpd_18	−1.16	1.08	1.95	1.49	4
cpd_19	−0.6	−0.72	−0.49	−0.64	4
cpd_20	2.47	−0.76	−0.49	−0.64	4

[a] Activity class: 1 (most potent), 2 (moderately potent), 3 (slightly active), and 4 (inactive).

Let λ be set as 0.01; separate OLR of the quality on each standardized predictor yields the coefficients a_{1j} of π^*, E_s^*, I^*, and X^* equal to, respectively, 0.5183 (−26.57), −1.1304 (−24.17), 0.0209 (−27.42), and 0.3907 (−26.96), with function g given in parentheses. The p values yielded by the Wald test on the four OLR coefficients are, respectivey, 0.22, 0.02, 0.96, and 0.36. Only π^* and E_s^* are significant variables at the 25% risk level.

After normalizing the coefficients, the first latent variable is defined as

$$\mathbf{t}_1 = \frac{0.5183 \times \pi^* + (-1.1304) \times E_s^*}{\sqrt{(0.5183)^2 + (-1.1304)^2}} = 0.4168\pi^* - 0.9090E_s^* = \begin{pmatrix} -0.73 \\ -1.31 \\ \vdots \\ 1.72 \end{pmatrix}$$

In order to obtain the second latent variable $\mathbf{t}_2$, $\mathbf{t}_2$ is built starting from the residuals $\mathbf{x}_{1j}$ of the regressions of each $\mathbf{x}_j$ on $\mathbf{t}_1$. In the case of $\mathbf{x}_{11}(\pi_1^*)$, for instance,

$$\mathbf{x}_{11} = 0.5497\mathbf{t}_1 + \text{residual}$$

$$\mathbf{x}_{21} = \mathbf{x}_{11} - 0.5497\mathbf{t}_1 = \begin{pmatrix} -0.69 \\ -0.79 \\ \vdots \\ 2.47 \end{pmatrix} - 0.5497\begin{pmatrix} -0.73 \\ -1.31 \\ \vdots \\ 1.72 \end{pmatrix} = \begin{pmatrix} -0.29 \\ -0.07 \\ \vdots \\ 1.53 \end{pmatrix}$$

$\mathbf{X}_2$ is obtained from the similar calculation

$$\mathbf{X}_2 = \begin{pmatrix} -0.29 & -0.13 & 1.77 & -0.71 \\ -0.07 & -0.03 & -0.81 & 1.35 \\ \vdots & \vdots & \vdots & \vdots \\ 1.53 & 0.70 & -0.07 & -0.46 \end{pmatrix}$$

Successively, separate OLR of the quality on each standardized predictor yields the coefficients a_{2j} of π_2^*, E_{s2}^*, I_2^* and X_2^* equal to, respectively, −0.0382 (−23.66), −0.0382 (−23.66), 0.4225 (−23.24), and 0.5176 (−23.00), with function g given in parentheses. The follwing p values are obtained for the predictors' coefficients: 0.93, 0.93, 0.39, and 0.28. There is no significant variable at the 25% risk level; the model with only the first latent variable $\mathbf{t}_1$ is retained.

Each activity class probability is estimated in the following:

$$\text{Prob}(\text{Activity} = 1) = 1 + \exp(-1.8903 - 1.3696\mathbf{t}_1)\}^{-1}$$

$$\text{Prob}(\text{Activity} \leq 2) = \{1 + \exp(-0.8167 - 1.3696\mathbf{t}_1)\}^{-1}$$

$$\text{Prob}(\text{Activity} \leq 3) = \{1 + \exp(1.0093 - 1.3696\mathbf{t}_1)\}^{-1}$$

By expressing $\mathbf{t}_1$ in terms of the normalized variables π^* and E_s^*,

$$\text{Prob}(\text{Activity} = 1) = \{1 + \exp(-1.8903 - 0.5708\pi^* + 1.2450E_s^*)\}^{-1}$$

$$\text{Prob}(\text{Activity} \leq 2) = \{1 + \exp(-0.8167 - 0.5708\pi^* + 1.2450E_s^*)\}^{-1}$$

$$\text{Prob}(\text{Activity} \leq 3) = \{1 + \exp(1.00903 - 0.5708\pi^* + 1.2450E_s^*)\}^{-1}$$

Supporting Information Available: List of inhibitory activities and SMILES keys of AChE data set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *197*, 178–180.

(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) Ohgaru, T.; Shimizu, R.; Okamoto, K.; Kawase, M.; Shirakuni, Y.; Nishikiori, R.; Takagi, T. Ordinal Classification Using Comparative Molecular Field Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 207–212.

(4) Bryson, M. C.; Johnson, M. E. The Incidence of Monotone Likelihood in the Cox Model. *Technometrics* **1981**, *23*, 381–383.

(5) Albert, A.; Anderson, J. A. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* **1984**, *71*, 1–10.

(6) Heinze, G.; Schemper, M. A Solution to the Problem of Separation in Logistic Regression. *Stat. Med.* **2002**, *21*, 2409–2419.

(7) Fort, G.; Lacroix, S. L. Classification Using Partial Least Squares with Penalized Logistic Regression. *Bioinformatics* **2005**, *21*, 1104–1111.

(8) Cessie, S. L.; Van Houwelingen, J. C. Ridge Estimators in Logistic Regression. *Appl. Stat.* **1992**, *41*, 191–201.

(9) Ding, B. Y.; Gentleman, R. Classification using generalized partial least squares. http://www.bepress.com/cgi/viewcontent.cgi?article=1004&context=bioconductor (accessed Feb 21, 2006).

(10) Bastein, P.; Vinzi, V. E.; Tenenhaus, M. PLS generalised linear regression. *Comput. Stat. Data Anal.* **2005**, *48*, 17–46.

(11) Coats, E. A. The CoMFA Steroids as A Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discovery Des.* **1998**, *12/13/14*, 199–213.

(12) Liu, S. S.; Yin, C. S.; Li, Z. L.; Cai, S. X. QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321–329.

(13) Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A. GRID Formalism for the Comparative Molecular Surface Analysis: Application to the CoMFA Benchmark Steroids, Azo Dyes, and HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1423–1435.

(14) Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. Improving the Performance of SOMFA by use of Standard Multivariate Methods. *SAR QSAR Environ. Res.* **2005**, *16*, 567–579.

(15) Castillo-Garit, J. A.; Ponce, Y. M.; Torrens, F. Atom-based 3D-Chiral Quadratic Indices. Part 2: Prediction of the Corticosteroid-Binding Globulinbinding Affinity of the 31 Benchmark Steroids Data Set. *Bio. Med. Chem.* **2006**, *14*, 2398–2408.

(16) Gasteiger, J. 31 Steroids Binding to the Corticosteroid Binding Globulin (CBG) Receptor. http://www2.chemie.uni-erlangen.de/services/steroids/index.html (accessed Feb 21, 2006).

(17) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

(18) Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **1993**, *80*, 27–38.

(19) Claeskens, G.; Croux, C.; Kerckhoven, J. V. Variable Selection for Logistic Regression Using a Prediction-Focused Information Criterion. *Biometrics* **2006**, *62*, 972–979.

(20) Klieber, M. A.; Underhill, C.; Hammond, G. L.; Muller, Y. A. Corticosteroid-Binding Globulin, A Structural Basis for Steroid Transport and Proteinase-Triggered Release. *J. Biol. Chem.* **2007**, *282*, 29594–29603.

(21) Kryger, G.; Silman, I.; Sussman, J. L. Three-dimensional Structure of a Complex of E2020 with Acetylcholinesterase from Torpedo californica. *Structure* **1999**, *7*, 297–307.

(22) Gieleciak, R.; Polanski, J. Modeling Robust QSAR. 2. Interative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pKa Values. *J. Chem. Inf. Model.* **2007**, *47*, 547–556.

(23) Martin, Y. C.; Holland, J. B.; Jarboe, C. H.; Plotnikoff, N. Discriminant Analysis of the Relationship between Physical Properties and the Inhibition of Monoamine Oxidase by Aminotetralins and Aminoindans. *J. Med. Chem.* **1974**, *17*, 409–413.

CI700444Z