Ordinal Classification Using Comparative Molecular Field Analysis

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

Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871, Japan, Tanabe Seiyaku Co., Ltd., 3-16-89 Kashima, Yodogawa-Ku, Osaka, 532-8505, Japan, Faculty of Pharmacy, Osaka Ohtani University, 3-11-1 Nishikiorikita, Tondabayashi, Osaka, 584-8540, Japan, and Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka, 565-0871, Osaka, Japan

Received July 4, 2007

Comparative Molecular Field Analysis (CoMFA) is most widely used as one of the 3-dimensional QSAR (3D-QSAR) methods to identify the relationship between chemical structure and biological activity. Conventional 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 a rating scale. Hence, rating classification-oriented CoMFA coupled with ordinal logistic regression has been developed, and its predictive ability and 3D graphical analysis ability have been investigated. As a result, this novel CoMFA (Logistic CoMFA) has been found to be more robust than conventional CoMFAs in both predictive and 3D graphical analysis abilities. Furthermore, Logistic CoMFA is useful since it can provide the probability of each rank.

INTRODUCTION

A detailed understanding of the quantitative structure−activity relationship (QSAR) is one of the principal goals of medicinal chemistry. To be able to clarify the relationship between chemical structure and biological activity is very important, particularly in the hit-to-lead stage of drug discovery. Researchers need to identify various properties of a large number of compounds in a limited period of the hit-to-lead stage. The growing need for early ADMET[1−3] increases the number of biological assays, such as Caco-2 cell permeability, CYP families inhibition, and hERG blockade, per compound. Unfortunately there are more experimental errors in screening data in the early screening stage of drug discovery than in reliable assays employed in the late stage. Since QSAR analysis generally makes use of IC_{50} and pK_i values as the indices of biological response, non-negligible differences between experimental and true IC_{50}/pK_i values can be found in some screening assays.[4,5] In addition, there are many in vivo assays where biological activity is measured only by a rating scale. These circumstances make it difficult to build a good QSAR model.

Prediction of activity rating, in which the potency of a compound is rather roughly assigned, enables us to quantitatively analyze the data set, which has not been able to be quantitatively analyzed because of noise. Treatment of a couple of data would be necessary to determine the rating classification since the ratings are not expressed in a metric scale.

Several studies on the application of the rating classification to classical QSAR have been performed. Martin et al.[6] conducted a classical QSAR analysis of monoamine oxidase inhibition by using a rating scale with linear discriminant analysis (LDA). Dunn, W. J., III et al.[7] analyzed the QSAR of β-adrenergic agents with the SIMCA (Soft Independent Modelling of Class Analogy) method,[8] which is based on a pattern recognition technique. LDA and SIMCA methods are not considered to be suitable for rating classification because both are introduced under the assumption that classes are independent. To harness the characteristics of ordinal classes, Takahashi et al.[9] developed ORMUCS (ORdered Multicategorical Classification using Simplex optimization technique). ORMUCS is also a pattern recognition method that determines a discriminant function using a simplex optimization. Apart from these methods, it is possible to apply the ordinal logistic regression method (OLR) to QSAR analysis. OLR is considered a statistical method that uses the probability of each rating for classification. In fact, OLR is one of the most popular methods used in social psychological studies and is more often applied to clinical data.[10,11]

Comparative Molecular Field Analysis (CoMFA) has become one of the most widely used 3-dimensional QSAR (3D-QSAR) methods[12−14] since it was introduced by Cramer et al.[15] to identify the relationship between 3-dimensional molecular structure and biological activity. Prevalence of commercial cheminformatics tools, such as Sybyl and high-performance CPU, makes it convenient to use 3D-QSAR analysis. However, the rating classification-oriented 3D-QSAR method has not yet been developed. Considering the prevalence of 3D-QSAR, it is desirable to classify rating with a 3D-QSAR method. 3D-QSAR analysis with SIMCA has not been used for rating classification and, unfortunately, has been limited to dichotomous (active/inactive) analysis[16] or selectivity analysis.[17]

In this study, we present the development and applicability of a novel rating classification-oriented CoMFA with OLR and compare it to conventional CoMFA analysis using 2 data sets. One data set is the corticosteroid binding globulin

* Corresponding author e-mail: satan@gen-info.osaka-u.ac.jp.
† Graduate School of Pharmaceutical Sciences, Osaka University.
‡ Tanabe Seiyaku Co., Ltd.
§ Osaka Ohtani University.
|| Research Institute for Microbial Diseases, Osaka University.

10.1021/ci700238k CCC: $40.75
Published on Web 12/28/2007

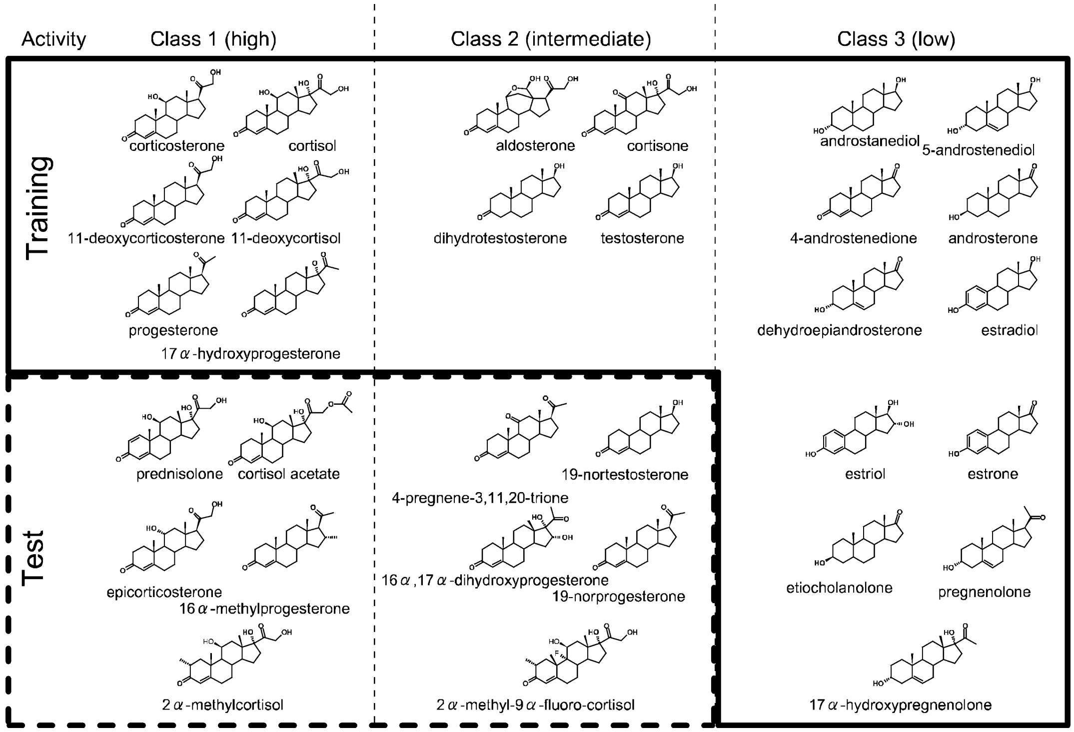

Figure 1. Composition and structure of steroids in the CBG data set.

(CBG) receptor binding data set, and it is popular for its 3D-QSAR benchmark. The other data set is the angiotensin converting enzyme (ACE) inhibitor data set. Each data set will be analyzed from 2 aspects of CoMFA method. First, the prediction accuracy of novel CoMFA will be investigated. Next, contour map analysis will be performed to clarify the portions important for keeping interaction between protein and ligand.

METHODS

1. Data Sets. *1.1. CBG Data Set.* Activities and chemical structures of the CBG data set ligands were used for validation of novel CoMFA-based analysis. Some groups have reported validation of each original 3D-QSAR using the CBG data set as a 3D-QSAR benchmark.[18−20] The CBG data set comprises 21 compounds for training and 10 compounds for test[15] (Figure 1). From the Web site,[21] not only activity values but also rating classes of the CBG data set are obtained (pIC_{50} values range widely from 5.00 to 7.88: Class1 ($pIC_{50} > 7.0$), Class2 ($5.8 < pIC_{50} \leq 7.0$), and Class3 ($pIC_{50} \leq 5.8$)). Unfortunately, there is no steroid of Class3 in the test set. Of course, it is desirable that all rating classes are included in the test set. However, in this study we used the data set without modifications, such as shuffling between the training set and the test set, since the CBG data set is a 3D-QSAR benchmark. The 3D coordinates were calculated by the 3D structure generator CORINA by Dr. Johann Gasteiger's group.[21]

1.2. ACE Data Set. Activities and chemical structures of the ACE data set ligands were also used for validation of novel CoMFA-based analysis. The ACE data set consists of a series of 31 inhibitors (22 training compounds and 9 test compounds) selected from DePriest's report,[22] which describes the use of the ACE data set for CoMFA modeling. In this study, we used 3D coordinates and partial charges as described by Sutherland et al.[23] (pIC_{50} values range widely from 2.74 to 8.96, and activity classes are allocated as follows: Class1 ($pIC_{50} > 7.2$), Class2 ($4.0 < pIC_{50} \leq 7.2$), and Class3 ($pIC_{50} < 4.0$) (Figure 2)). Unlike the CBG data set, the test set of the ACE data set comprises all activity classes.

2. Molecular Modeling. The 3D coordinates were used without any refinement for both data sets. Gasteiger−Hückel charges were assigned to each atom by Sybyl versions 7.22 (Tripos Inc.) only for the CBG data set. For the ACE data set, atomic partial charges in mol2 files obtained from Sutherland's study[23] were used.

3. Calculation of Steric and Electrostatic Potential Fields. The CoMFA methodology of 3D-QSAR is based on the assumption that interactions between a ligand and its receptor are primarily noncovalent in nature and shape-dependent. Therefore, QSAR can be derived by sampling the steric and electrostatic fields surrounding a set of ligands and correlating the differences in those fields to biological activity. 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 (Tripos standard field, dielectric constant $1/r$, cutoff 30 kcal/mol, a volume dependent lattice with 2 Å step size in each direction beyond the aligned molecules) using an sp^3 carbon probe atom with a charge of $+1.0$. The CoMFA lattice for the CBG series was $9 \times 11 \times 9$ Å ($X = -8.42$ to 7.58, $Y = -4.23$ to 15.77, $Z = -10.60$ to 5.40) with 891 points, and the CoMFA lattice for the ACE series was $11 \times 11 \times 10$ Å ($X = -9.58$ to 12.19, $Y = -14.92$ to 6.37, $Z = -7.70$ to 10.97) with 1210 points.

4. Comparison of Novel CoMFA with Several Conventional CoMFAs. The Partial Least-Squares (PLS) or the

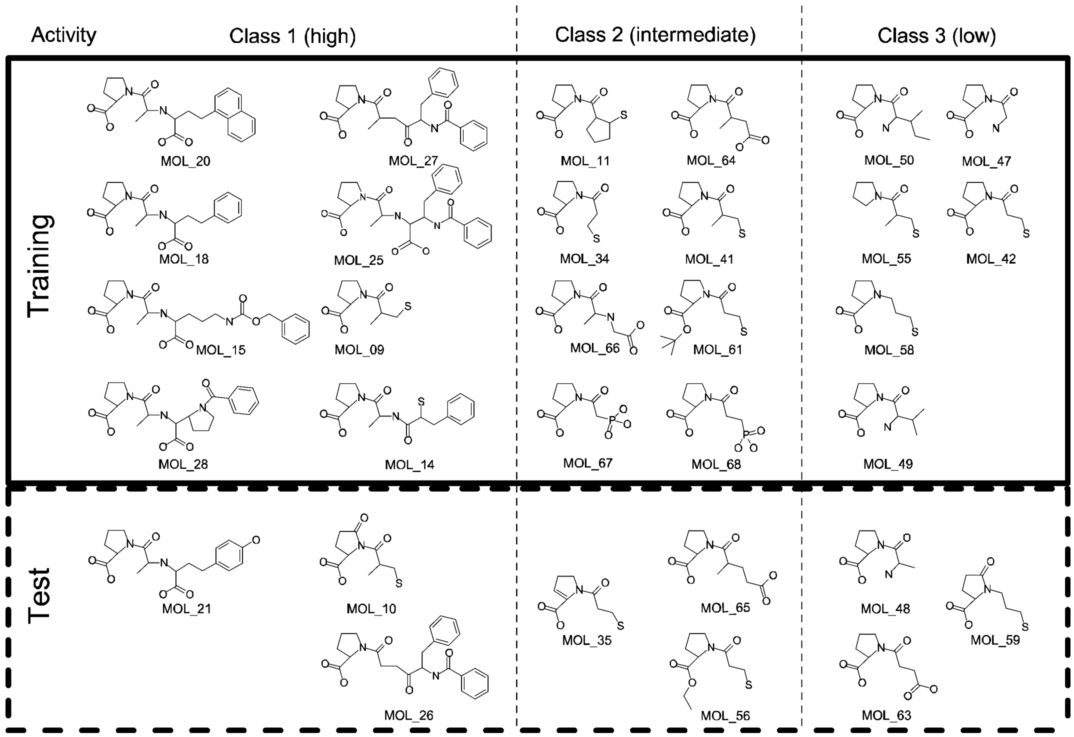

Figure 2. Composition and structure of pyrrolidine series in the ACE data set.

Principal Component Regression (PCR) method is used in the process of conventional CoMFA. In this study, only the PLS method was used to perform novel CoMFA as well as conventional CoMFA.

Ordinal logistic regression analysis was coupled with CoMFA-PLS in this novel CoMFA. Basically, logistic regression analysis is chosen for analyzing dichotomous data. Using 2 logistic functions enables us to analyze the rating classification. Latent variables calculated by PLS were applied as explanatory variables of ordinal logistic regression. The following shows details of the novel CoMFA (Logistic CoMFA).

Logistic CoMFA: OLR analysis gives the probability of each rank. The data set is categorized into 3 rating classes, and the respective probabilities are

$$\mathrm{Prb(Class1)} = \{1 + \exp(-\eta_1)\}^{-1} \tag{1}$$

$$\mathrm{Prb(Class2)} = \{1 + \exp(-\eta_2)\}^{-1} - \{1 + \exp(-\eta_1)\}^{-1} \tag{2}$$

$$\mathrm{Prb(Class3)} = 1 - \{1 + \exp(-\eta_2)\}^{-1} \tag{3}$$

where η_1 and η_2 are rewritten as

$$\eta_1 = \alpha_1 - \beta'\mathbf{t} \tag{4}$$

$$\eta_2 = \alpha_2 - \beta'\mathbf{t} \tag{5}$$

$$\alpha_1 \leq \alpha_2 \tag{6}$$

where $\mathbf{t}$ is a set of latent variables, which are introduced from steric and electrostatic potentials (explanatory variables) $\mathbf{X}$ by PLS. Figure 3 shows curves of the cumulative probabilities Prb(Class1) and Prb(Class1 or Class2). Latent variables are extracted as long as the leave-one-out (LOO) cross-validated Spearman's rank coefficient (q_S) increases.

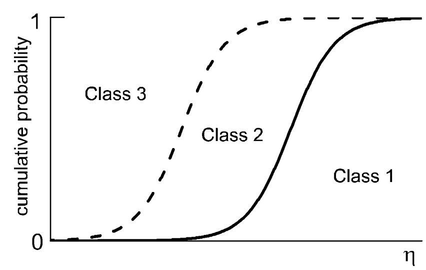

Figure 3. Cumulative probability of each class vs η.

The coefficients α_1, α_2, and β are evaluated by maximum likelihood estimation (MLE). In general, MLE yields values for unknown parameters, which maximize the probability of the observed set of data. The conjugate-gradient numerical optimization algorithm is adopted to maximize the log-likelihood.[24]

In contrast, 3 types of conventional CoMFAs were used to examine the prediction accuracy of rating classification. The methods are described below.

CoMFA1: Observed pK_i values are learned and expected pK_i values are calculated. Expected K_i classes are then determined using the expected pK_i values.

CoMFA2: Observed K_i classes are learned without any treatment and expected K_i classes are calculated.

CoMFA3: Scales between classes are not necessarily equivalent, in which case, it is often effective to use the average rank from the top toward each rating class. In CoMFA3, average ranks are learned and expected ranks are calculated. Expected K_i classes are then determined using expected average ranks.

5. CoMFA Program. CoMFA1 was carried out with Sybyl version 7.22. Logistic CoMFA and other conventional CoMFAs (CoMFA2 and CoMFA3), programmed in Fortran90, were computed on a dual core Xeon 2.0 GHz computer.

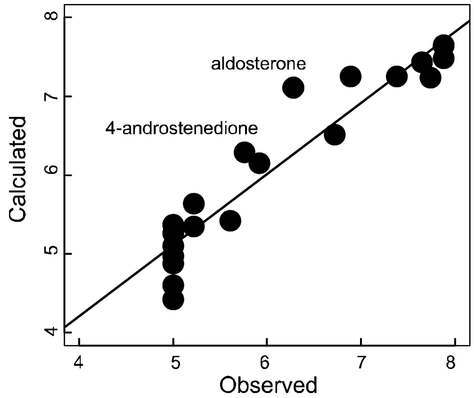

Figure 4. Predicted vs observed pK_i of the 21 training steroids.

Table 1. Summary of Leave-One-Out Cross-Validation of the CBG Training Set by Logistic and Conventional CoMFA Analyses

	CoMFA		
	Logistic	CoMFA2	CoMFA3
q_S[a]	0.75	0.75	0.81
no. of cmpnts	1	4	5
no. of correct	13	17	17
accuracy	62%	81%	81%

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

Table 2. Prediction of the CBG Test Set by Logistic and Conventional CoMFA Analyses

	CoMFA		
	Logistic	CoMFA2	CoMFA3
q_S[a]	0.45	0.077	−0.024
no.of correct	7	6	4
accuracy	70%	60%	40%

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

RESULTS AND DISCUSSION

1. Steroids of CBG Binding Analysis. *1.1. Validation of the CBG Data Set.* Figure 4 shows the CoMFA1 model of the CBG data set (number of latent variables = 2, $r^2 = 0.901$, cross-validated $r^2 = 0.804$). Almost all compounds satisfied the CoMFA1 model. Although predicted pK_is of aldosterone and 4-androstenedione were distant from observed pK_is, the CoMFA1 model, as a whole, can elucidate the relationship between structure and activity. The results also indicate that CBG activity was measured with relatively high precision.

Next, Logistic CoMFA, CoMFA2, and CoMFA3 models were performed with LOO cross-validation (Table 1). All models were good in terms of q_S, but CoMFA3 was found to be the best model to calculate accurate activity ratings. All models were accurately analyzed using 1−5 latent variables. The fact that the best number of latent variables is small means that the model is simple and is expected to readily interpret contour maps.

Figure 5. Profiles of each class probability vs η.

The prediction capability of each model was next investigated using the best numbers of latent variables. As a result, Logistic CoMFA prediction was found to be the most accurate of all. CoMFA3 was found to be the worst model in terms of prediction accuracy as well as q_S because of overfitting (Table 2). CoMFA2 and CoMFA3 use more latent variables than Logistic CoMFA to obtain the best PRESS values. Besides, considering Logistic CoMFA, the probability of each class, shown in Figure 5, changes gradually vs η. In this case, the pK_i class changes from 1 to 3 with decreased η. Thus, Logistic CoMFA, unlike CoMFA2 and CoMFA3, can give the probability of each rank (Table 3).

1.2. Contour Interpretation. Logistic CoMFA was found to be more robust than conventional CoMFA methods as it is required to predict portions of the molecule to improve the binding affinity. The contour maps obtained by CoMFA show how 3D-QSAR methods are useful to identify features important 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 6 shows a comparison of contour maps (STDEV*Coeff) derived from the CoMFA1 model and the Logistic CoMFA model. The cortisol centered molecule is strongly bound to CBG, and reduction of C-3 causes low binding affinity. This is supported by both contour maps. In addition, the CoMFA1 map supports the fact that the carbonyl at position 17 causes low activity, though the Logistic CoMFA map, unfortunately, does not support this idea. This is probably because the numbers of latent variables used are distinct between Logistic CoMFA and CoMFA1. Each latent variable holds information of portions important for activity. In this study, Logistic CoMFA uses only one latent variable, which probably contains the portions around C-3 and C-11. Interestingly,

Table 3. Prediction and Probability of Activity Classes of the 10 Test Steroids by Logistic CoMFA

steroids	$Class_{obsd}$	$Class_{pred}$	Prb(Class1)	Prb(Class2)	Prb(Class3)
16α,17α-dihydroxyprogesterone	2	2	0.46	0.47	0.07
16α-methylprogesterone	1	1	0.81	0.17	0.02
19-norprogesterone	2	2	0.34	0.55	0.11
19-nortestosterone	2	3	0.01	0.12	0.87
2α-methyl-9α-fluorocortisol	2	1	0.99	0.01	0.00
2α-methylcortisol	1	1	0.99	0.01	0.00
4-pregnene-3,11,20-trione	2	2	0.26	0.58	0.16
cortisol acetate	1	3	0.00	0.00	1.00
epicorticosterone	1	1	0.72	0.26	0.02
predonisolone	1	1	0.90	0.09	0.01

Figure 6. Comparison of contour maps (STDEV*Coeff) derived from the conventional CoMFA model (left) and the classification-oriented Logistic CoMFA model (right). The centered molecule, cortisol, is strongly bound to CBG. For an increase of the activity, the positive charge in blue and the negative charge in red have to be increased. The molecular volume has to be increased (green) or decreased (yellow) to increase the activity.

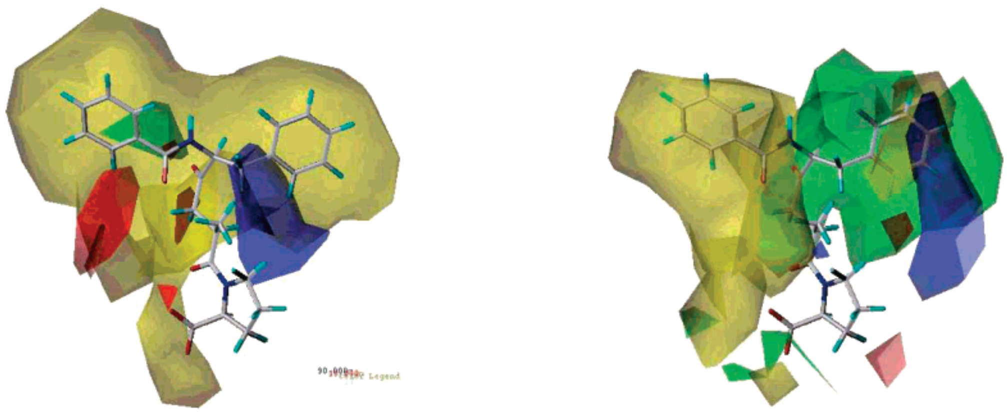

Figure 7. Comparison of contour maps (STDEV*Coeff) derived from the conventional CoMFA model (left) and the classification-oriented Logistic CoMFA model (right). The centered molecule, MOL_27, is strongly bound to CBG. For an increase of the activity, the positive charge in blue and the negative charge in red have to be increased. The molecular volume has to be increased (green) or decreased (yellow) to increase the activity.

Table 4. Summary of Leave-One-Out Cross-Validation of the ACE Data Set by Logistic and Conventional CoMFA Analyses

	CoMFA		
	Logistic	CoMFA2	CoMFA3
q_S[a]	0.69	0.67	0.67
no. of cmpnts	3	3	3
no. of correct	13	14	14
accuracy	59%	64%	64%

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

Table 5. Prediction of the ACE Test Set by Logistic and Conventional CoMFA Analyses

	CoMFA		
	Logistic	CoMFA2	CoMFA3
q_S[a]	0.62	0.54	0.54
no. of correct	7	7	7
accuracy	78%	78%	78%

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

hydroxylation of C-11, which is held by Logistic CoMFA but not by CoMFA1, leads to high activity. Therefore, in order to obtain an ideal contour map, we need to improve the selection of latent variables. The result indicates that important information is retained by use of ordinal classes although CoMFA1 uses pK_i values, while Logistic CoMFA uses pK_i classes, that is, rounded pK_i.

2. ACE Inhibitors. *2.1. Validation of the ACE Data Set.* Shown in Table 4 are the results from LOO cross-validation of each CoMFA method. Logistic CoMFA was found to be the best model in terms of both q_S. As for accuracy rate, CoMFA2 and CoMFA3 were found to be slightly better than Logistic CoMFA. Thereafter, the prediction capacity of each model was investigated using the best numbers of latent variables. As it turned out, the Logistic CoMFA prediction was merely found to be as accurate as CoMFA2 or CoMFA3. Still, from the q_S point of view, Logistic CoMFA marked the best value of the 3 models. Thus, Logistic CoMFA can be considered as the most robust method to perform rating classification analysis.

2.2. Contour Interpretation. Contour maps of CoMFA1 and Logistic CoMFA are shown in Figure 7 with MOL_27 depicted in the center. Both maps exhibit steric potential, which is more important than electrostatic potential for inhibitory activity. The CoMFA1 contour map clearly shows a narrow space behind the benzamido group and the benzyl group. The Logistic CoMFA map also shows a narrow space around the benzamido group and behind the benzyl group. Furthermore, the Logistic CoMFA map clearly exhibits a free space in front of the benzyl group. This emphasizes the fact that MOL_25 and MOL_27 are stronger inhibitors. Interestingly, in some cases Logistic CoMFA makes it possible to grasp the substituent effect better than CoMFA1. Thus, ordinal classification of the activity can facilitate understanding of the structure−activity relationship.

CONCLUSION

In the present study, we compared the performance of different classification methods based on CoMFA analyses. Logistic CoMFA, which couples OLR with CoMFA, was found to be the most robust model with respect to not only prediction accuracy but also graphical analysis. Moreover, Logistic CoMFA, unlike conventional CoMFAs, has been shown to be a statistical method that gives the probability of each rank. As a great amount of rank-scale biological activity data are produced in the process of drug discovery, we believe that Logistic CoMFA is a novel solution to analyze rating data and to facilitate novel drug developments.

ACKNOWLEDGMENT

We thank Prof. G. Marshall for providing us with the ACE data set and Dr. D. Weaver for correcting part of the ACE data set. We are also grateful to Drs. Ohmizu, Ogiku, Fukushima, Nakao, and Kuroda (Mitsubishi Tanabe Pharma Corporation) for their helpful advice and comments on QSAR.

REFERENCES AND NOTES

(1) Fujikawa, M.; Ano, R. . et al. Relationships between Structure and High-Throughput Screening Permeability of Diverse Drugs with Artificial Membranes: Application to Prediction of Caco-2 Cell Permeability. *Bio. Med. Chem*. **2005**, *13*, 4721−4732.

(2) Jenkins, K. M.; Angeles, R. . et al. Automated High Throughput ADME Assays for Metabolic Stability and Cytochrome P450 Inhibition Profiling of Combinatorial Libraries. *J. Pharm. Biomed. Anal.* **2004**, *34*, 989−1004.

(3) Deacon, M.; Singleton, D. . et al. Early Evaluation of Compound QT Prolongation Effects: A Predictive 384-well Fluorescence Polarization Binding Assay for Measuring hERG Blockade. *J. Pharmacol. Toxicol. Methods* **2007**, *55*, 238−247.

(4) Agarwal, A.; Pearson, P. P. . et al. Three-dimensional Quantitative Structure-Activity Relationships of 5-HT Receptor Binding Data for Tetrahydropyridinylindole Derivatives: A Comparison of the Hansh and CoMFA Methods. *J. Med. Chem*. **1993**, *36*, 4006−4014.

(5) Tsigelny, I.; Grant, B. D. . et al. Catalytic Subunit of cAMP-dependent Protein Kinase: Electrostatic Features and Peptide Recognition. *Biopolymers* **1996**, *39*, 353−365.

(6) Martin, Y. C.; Holland, J. B. . et al. 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.

(7) Dunn, W. J., III; Wold, S. . et al. Structure-Activity Study of β-adrenergic Agents Using the SIMCA Method of Pattern Recognition. *J. Med. Chem*. **1978**, *21*, 922−930.

(8) Wold, S. Pattern Recognition by Means of Disjoint Principal Components Models. *Pattern Recognit*. **1976**, *8*, 127−139.

(9) Takahashi, Y.; Miyashita, Y. . et al. A New Approach for Ordered Multicategorical Classification Using Simplex Technique. *Bunseki Kagaku* **1984**, *33*, 487−494.

(10) Antill, Y.; Reynolds. J. . et al. Risk-Reducing Surgery in Women with Familial Susceptibility for Breast and/or Ovarian Cancer. *Eur. J. Cancer*. **2006**, *42*, 621−628.

(11) Mukherjee, B.; Liu, I. et al. Analysis of Matched Case-control Data with Multiple Ordered Disease States: Possible Choices and Comparisons. *Statistics Med.* **2007**, *26*, 3240−3257.

(12) Nakao, K.; Asao, M. . et al. Benzoxazoline and Benzimidazoline Derivatives as Novel Aldose Reductase Inhibitors, Part 1: Lead Evolution. *Med. Chem. Res*. **1999**, *9:7/8*, 621−630.

(13) Nakao, K.; Asao, M. . et al. Benzoxazoline and Benzimidazoline Derivatives as Novel Aldose Reductase Inhibitors, Part 2: Lead Optimization. *Med. Chem. Res*. **1999**, *9:7/8*, 631−642.

(14) Nakao, K.; Asao, M. . et al. 3D-pharmacophore Analyses of Aldose Reductase Inhibitory Spiroquinazolinones. *Drug Des. Discovery* **1999**, *16*, 155−163.

(15) Cramer, R. D.; Patterson, D. E. . et al. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.

(16) Merino, V. M.; Cerecetto, H. CoMFA-SIMCA Model for Antichagasic Nitrofurazone Derivatives. *Bio. Med. Chem*. **2001**, *9*, 1025−1030.

(17) Sutherland, J. J.; Weaver, D. F. Three-dimensional Quantitative Structure-activity and Structure-Selectivity Relationships of Dihydrofolate Reductase Inhibitors. *J. Comput.-Aided Mol. Des*. **2004**, *18*, 309−331.

(18) Liu, S.-S.; Yin, C.-S. . et al. QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 321−329.

(19) Polanski, J.; Gieleciak, R. . et al. 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.

(20) Korhonen, S.-P.; Tuppurainen, K. . et al. Improving the Performance of SOMFA by Use of Standard Multivariate Methods. *SAR QSAR Environ. Res.* **2005**, *16*, 567−579.

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

(22) DePriest, S. A.; Mayer, D. . et al. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc*. **1993**, *115*, 5372−5384.

(23) Sutherland, J. J.; O'Brien L. A. . et al. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(24) Fok, D.; Franses, P. H. Ordered Logit Analysis for Selectively Sampled Data. *Comp. Stat. Data Anal.* **2002**, *40*, 477−497.

CI700238K